ESD-TR-79-154

AD A077355

(12)

LEVEL#

SC

Technical Report

535

Digital Encoding of Speech and Audio Signals Based on the Perceptual Requirements of the Auditory System

M. A. Krasner

18 June 1979

Prepared for the Defense Advanced Research Projects Agency
under Electronic Systems Division Contract F19628-78-C-0002 by

Lincoln Laboratory
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LEXINGTON, MASSACHUSETTS

Approved for public release; distribution unlimited.

Del 1473

The work reported in this document was performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract F19628-78-C-0002 (ARPA Order 2006).

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Raymond L. Loiselle

Raymond L. Loiselle, Lt. Col., USAF
Chief, ESD Lincoln Laboratory Project Office

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LINCOLN LABORATORY

DIGITAL ENCODING OF SPEECH AND AUDIO SIGNALS BASED ON THE PERCEPTUAL REQUIREMENTS OF THE AUDITORY SYSTEM

M. A. KRASNER

Group 24

TECHNICAL REPORT 535

18 JUNE 1979

LEXINGTON MASSACHUSETTS

ABSTRACT

The development of a digital encoding system for speech and audio signals is described. The system is designed to exploit the limited detection ability of the auditory system. Existing digital encoders are examined. Relevant psychoacoustic experiments are reviewed. Where the literature is lacking, a simple masking experiment is performed and the results reported. The design of the encoding system and specifications of system parameters are then developed from the perceptual requirements and digital signal processing techniques.

The encoder is a multi-channel system, each channel approximately of critical bandwidth. The input signal is filtered via the quadrature mirror filter technique. An extensive development of this technique is presented. Channels are quantized with an adaptive PCM scheme.

The encoder is evaluated for speech and audio signal inputs. For 4.1-kHz bandwidth speech, the differential threshold of encoding degradation occurs at a bit rate of 34.4 kbps. At 16 kbps, the encoder produces toll-quality speech output. Audio signals of 15-kHz bandwidth can be encoded at 123.8 kbps without audible degradation.

This report is based on a thesis submitted to the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology on 4 May 1979 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

CONTENTS

ACKNOWLEDGMENTS

I would like to express my gratitude to Dr. Bernard Gold. His friendship, insights, and supervision made this research possible.

I would like to thank Prof. Kenneth N. Stevens and Prof. Henry J. Zimmermann for their advice and direction during the research.

I would also like to acknowledge Dr. Theodore Bially, Dr. Edward M. Hofstetter, Dr. Robert J. McAulay, Mr. Elliot Singer, Mr. Joseph Tierney, and Dr. Clifford J. Weinstein for invaluable discussions. Their willingness to comment on my research and writing was a major contribution.

DIGITAL ENCODING OF SPEECH AND AUDIO SIGNALS BASED ON THE PERCEPTUAL REQUIREMENTS OF THE AUDITORY SYSTEM

I. INTRODUCTION

Digital techniques for the processing, transmission, and storage of speech and other audio signals have become increasingly important in the past several years. Advantages of the digital domain include flexible processing not previously possible, increased transmission reliability, and error-resistant storage. Implicit in the conversion of a continuous audio waveform into a digital bit stream are degradations due to the nonlinearity of the process. While a sentence can be transmitted simply by coding exactly the letters of each word, there are an infinite number of waveforms that could represent that spoken sentence. Increasing the amount of information in the digital signal by increasing the digital bit rate can decrease the error in the digital approximation of the sentence waveform, but increase the costs of transmission and storage by necessitating the use of a channel of higher capacity.

Audio signals differ from other signals since the intent is to communicate with a person. The performance of an audio system can not be measured by a simple root-mean-square (RMS) error measurement. Rather, it is the complex processing of the auditory system that determines its quality. Indeed, an audio system that compares favorably to another system in a traditional signal-to-noise ratio (SNR) error measurement, the ratio of the mean square signal level to the mean square noise, may be judged as annoying to listen to and, therefore, of lower quality. An audio system with additive white noise with a SNR of 20 dB is generally preferred to a system with 10 percent harmonic distortion. For speech systems, even listener preference does not correlate well with intelligibility. For example, adding dither to a 3-bit-per-sample linear pulse code modulation (PCM) system does not affect the SNR of the system. The dithered system, however, is of lower intelligibility, but is preferred by listeners over the undithered PCM coder.[1] To design and evaluate an audio system, it is necessary to have some understanding of the functioning of the auditory system, its capabilities and limitations. With that understanding, it may be possible to identify subjective quality variables and relate them to objective physical quantities in the stimuli.

In this report, an encoding system is designed to exploit the limitation of the auditory system imposed by masking characteristics, the ability of one sound to inhibit the perception of another sound. The system is designed so that the error noise due to quantization is masked by the audio signal being encoded. The test of the system is whether a listener can perceive any differences between the original signal and the signal that has been processed by the encoding system.

A. Historical Development of the Problem

Digital encoding of audio signals can be grouped into three types of systems by their quality and applications. The highest quality coders have been developed for use with voice and music for the radio broadcast and record industries. These systems are characterized by wide signal bandwidths of 12 to 20 kHz and large SNRs of 50 to 100 dB. Bit rates of up to 500 kilobits per second (kbps) are common in these high-quality systems. The fidelity criterion for these systems is that the listener will perceive little or no degradation of the input signal.

Digital encoders for speech tend to emphasize lower bit rates since the objective is often a low-cost communication system as for telephone communications. Degradations are permitted as long as intelligibility is high and it is not annoying to listen to the sound for reasonable lengths of time. For commercial telephone applications, a 64-kbps system is commonly used. Although coders with comparable quality such as adaptive differential pulse code modulation (ADPCM)[2] have been developed using rates lower than 40 kbps (Chapter II-D-1), the implementation costs of these algorithms often outweigh the savings due to the use of lower capacity channels.

The third area of development has been in very low-rate speech communication systems needing channel capacities as low as 2.4 kbps. Invariably, this bit-rate reduction is achieved by modeling the production of the input speech by a slowly time-varying vocal-tract system. Information to specify the parameters of that production system are encoded. Although intelligibility is high for speech input with a low noise background, signals that do not fit the model, such as nonspeech and speech in a noisy environment are reproduced poorly.

For many applications, it is not yet economically feasible to use digital transmission and storage methods because of the cost of the high-capacity channels required. Encoding systems that would permit the use of lower capacity channels while maintaining the necessary signal quality would open new applications areas for digital techniques.

B. Scope of this Report

The objective of this report is to relate the results of psychoacoustic research to the development of a digital encoding system. By using the limitations of the auditory system, the system is made to be efficient, using only the bit rate necessary to maintain its quality.

The encoder is designed so that the degradations introduced through its processing are not audible when presented along with the audio signal. The encoder, based on the characteristics of the auditory system, should work well with speech, music, or any other audio signal.

The report is divided into several parts. Existing digital encoders are examined and relevant psychoacoustic experiments are reviewed. Where the literature is lacking, simple experiments are performed, and the results of these experiments are reported. The design of the system is then developed from the perceptual requirements and digital signal-processing techniques. The system is evaluated with high-quality speech and audio signals to determine parameters for broadcast-quality transmission and archival-quality storage. Experiments are performed to find the minimum bit rate such that the processing of the system is not noticeable to the average listener. The system parameters are then set for lower bit rates and the encoder compared to other encoders for possible use in basic speech communication systems.

II. REVIEW OF EXISTING DIGITAL ENCODERS

A. Introduction

Degradation of analog signal quality resulting from processing, storage, and transmission of speech and audio signals is often a major obstacle in the implementation of such systems. To alleviate this problem, digital techniques are being used increasingly for high-quality speech and audio systems. For voice communication over phone lines and satellite links, digital techniques simplify the multiplexing of several conversations and the protection from noise. As the technology has progressed, many algorithms for digital encoding have emerged.

Digitization requires two processes, sampling of the signal at discrete instants of time and quantization of the signal samples to a discrete number of bits of information. (This is not strictly true for some low-rate vocoders that try to model the speech production process. It is a valid assumption for the class of encoders and quantizers relevant to this research.) It is sometimes convenient to consider sampling and quantization to be separate processes even though they may be implemented together. The ordering of these processes is not important and is chosen to simplify the analysis in the Chapter.

For a band-limited signal, the process of sampling can be accomplished without any loss of information. By sampling at the Nyquist rate, a rate of twice the highest frequency present in the continuous-time signal, the sampling is a simply reversible process. Quantization, however, introduces error. It is the audibility of this error that encoding systems try to minimize.

For speech signals, voiced segments have very little energy above 4 kHz. Unvoiced speech sounds, however, have significant energy at frequencies greater than 8 kHz. Sampling at approximately 8 kHz for a resultant 4-kHz signal bandwidth is typical for telephone and similar communications. Very little loss of intelligibility is evidenced at that sampling rate. Larger bandwidths are used for higher-quality speech systems and for music. A frequency response to 15 kHz and higher is typical of broadcast quality audio encoders. Music can usually be filtered to 15 kHz with little or no audible degradation.

The most basic encoding system is pulse code modulation (PCM). Among its advantages are simplicity, direct representation as binary numbers for digital storage, and easy implementation of the corresponding analog-to-digital (A/D) and digital-to-analog (D/A) converters. Most other systems are derived from PCM.

B. Instantaneous Quantization

Instantaneous quantizers are characterized by memoryless input-output relations. The relation is nonlinear by necessity as the output is only permitted to take on a finite number of values.

In a continuous-time system, the output of a memoryless nonlinear process is periodic if the input is periodic. The output contains energy only at multiples of the fundamental frequency, the frequency of the periodicity. The error is harmonic distortion for sinusoidal inputs, and harmonic and intermodulation distortion for inputs that are sums of sinusoids. If a band-limited signal is quantized, the distortion products are not restricted to that band. When the resulting signal is sampled at a rate commensurate with the bandwidth of the unquantized signal, the components of the error signal outside of the original frequency band are aliased into that band at frequencies that are not necessarily related to the original signal. This error may then sound like white noise or harmonic distortion, depending on the exact quantization system used.

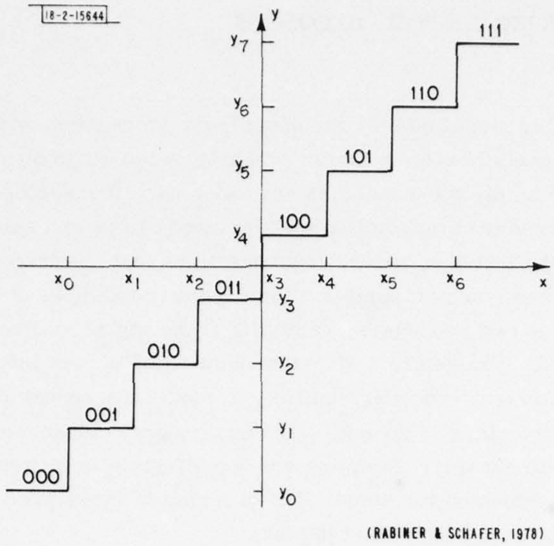


Fig. II-1. Linear quantization I/O characteristic.

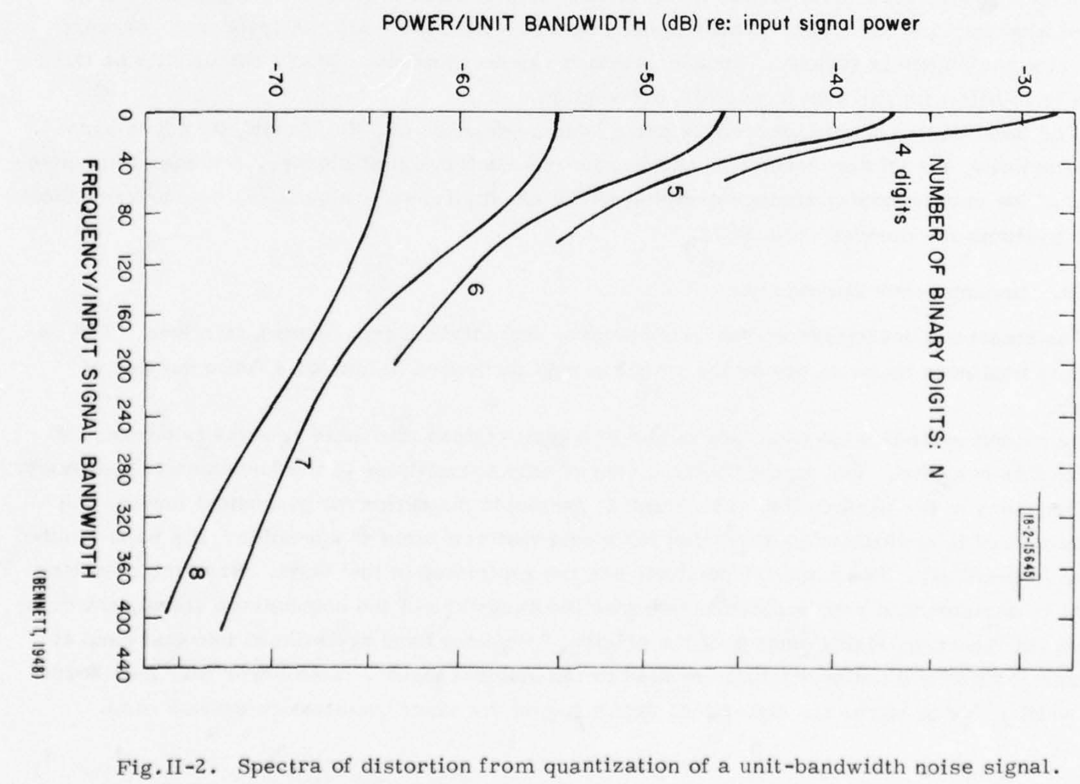


Fig. II-2. Spectra of distortion from quantization of a unit-bandwidth noise signal.

1. Pulse Code Modulation

PCM quantization divides the input amplitude range into a number of equally spaced intervals. The number of intervals is usually a power of 2 so that the interval into which a sample falls may be coded into an integer number of binary bits. A sample is approximated by the value in the middle of the interval into which it falls. This input-output relation is shown in Fig. II-1.

As the number of levels increase, the bandwidth of the quantized signal increases. Figure II-2 shows the bandwidth of distortion for the quantization of a unit bandwidth, unit-power random noise. Note that for 8-bit quantization (256 levels) the spectrum of the error has decreased only 10 dB in power from the low-frequency level by a frequency of 300 times the input bandwidth.[4] Using this graph, the perception of the quantization error for PCM can be explained.

One-bit PCM, 2 levels, is a hard-limiter transforming an input signal into a pulse train with time-varying duty cycle. Since the distortion products present in the hard-limited signal are predominantly within the original signal bandwidth, much of the error energy will not be aliased by sampling. The error is perceived as distortion of the input.

Ten-bit PCM, 1024 levels, represents a good approximation to the input. The error is very wideband as implied by Fig. II-2. When sampled, there is little correlation of the error with the input and it sounds like white noise.

As the number of bits increase, the level of the error decreases logarithmically. For an N-bit quantizer, the SNR for the maximum-level sinusoid input that will not overload the quantizer's range is in Eq. (II-1):

$$SNR = 6.02\ N + 1.76\ dB \quad . \tag{II-1}$$

If a signal such as speech is to be quantized, headroom must be left to prevent clipping on peaks larger than the average level. Assuming a Laplacian density for the amplitude of the speech samples, only 0.35 percent of the samples will be larger in magnitude than four standard deviations. Setting this equal to the maximum quantizer level, i.e., allowing for peak samples 12 dB greater than the RMS signal amplitude, the SNR is now 9 dB less than in Eq. (II-1), as in Eq. (II-2):[5]

$$SNR = 6.02\ N - 7.27\ dB \quad . \tag{II-2}$$

2. Dither

The error for a PCM quantization system is normally assumed to be statistically independent of the input waveform. This independence is caused by the aliasing of high-frequency components of the quantized signal into the signal band. For a small number of levels, this assumption is not valid. As discussed in Section II-B-1, the quantized and sampled signal will sound like the result of harmonic distortion when just a couple of bits are used. To remove the correlation of the error to the input, the scheme shown in Fig. II-3 can be used. White noise with a uniform probability density with width equal to one quantization interval is added to the signal before quantization. The same noise is subtracted after the quantizer. The dither noise has the effect of making the error be a white-noise process. In practice, the quantized signal, $y'(n)$, is transmitted or stored and it is not practical to provide the dither noise signal, $z(n)$, to the decoder. The dither noise is implemented as a pseudorandom sequence that can be generated by both the encoder and decoder without information exchange except for synchronization.

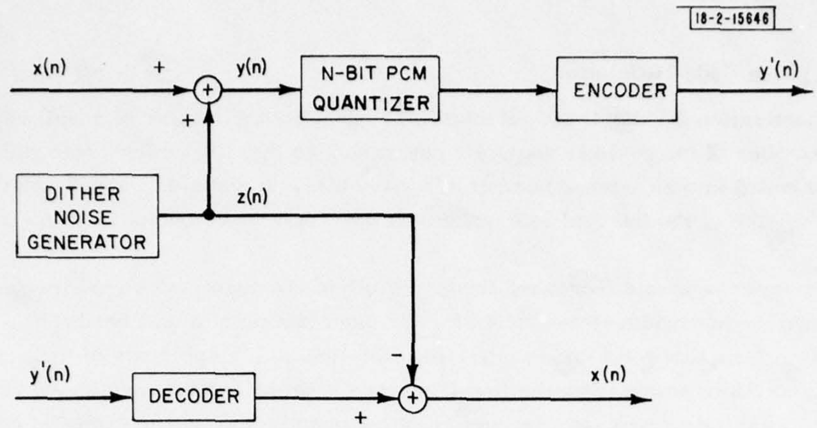


Fig. II-3. PCM quantizer with dither.

The error for the quantizer with dither can be shown to be zero mean white noise that is statistically independent of the input. The noise power is the same as it is for the system without dither.

Since harmonic distortion is not pleasant to listen to, the system with dither is rated as having a higher subjective quality than the PCM system without dither, even though the RMS error has not changed. It is interesting to note that adding dither to 2- and 3-bit PCM systems decreases the intelligibility while increasing subjective quality and listener preference.[1]

3. Instantaneous Companding

Most audio signals of interest vary in short-time average power over time. A PCM encoder, having equally spaced quantization intervals, will have an error that does not vary in amplitude for different input amplitudes. While the error may be tolerable for loud musical passages and speakers, it will be more audible for lower-volume time intervals. The signal power in speech may vary as much as 40 dB among speakers and environments. For example, a 7-bit PCM system set for a loud speech segment as in Section II-B-1 for full use of the quantizer's range would have an SNR of 35 dB. Another speech waveform might only use a few of the quantization levels and have an SNR of 10 dB or less. In general, an extra four bits are necessary in PCM to compensate for the wide variance of speech signal power. Music, where 60-dB differences of power in different passages are common, would require significantly more bits.

The problem of encoding signals with large dynamic range (the ratio of the largest and smallest short-time energy levels) can be reduced by companding. Companding is achieved by compression of the signal before quantization to reduce the dynamic range and subsequent expansion after quantization to undo the effects of compression. Companding is often combined with quantization by using a nonuniform distribution of the quantization levels in a PCM system. Thus it is often referred to as nonlinear PCM.

To maintain the SNR over any region of the input dynamic range, it is necessary to quantize the logarithm of the signal magnitude. Unfortunately, this requires an infinite number of quantization levels as the slope of the logarithm input-output relation is infinite for inputs near zero.

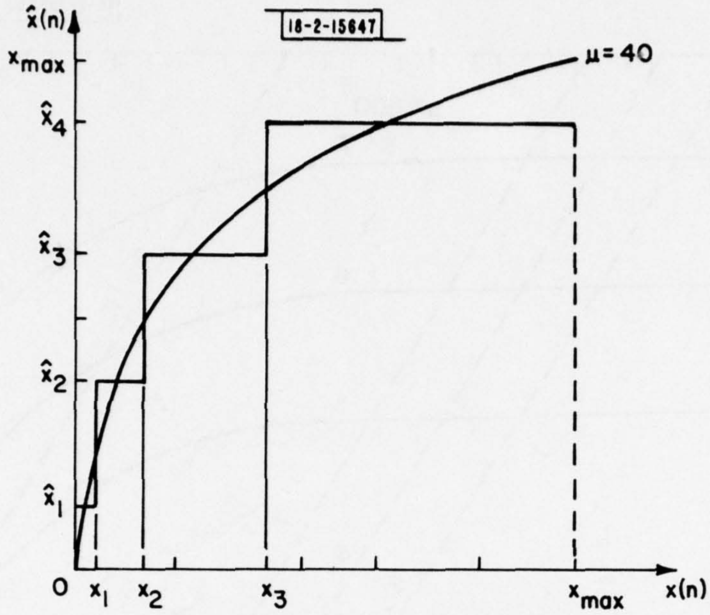


Fig. II-4. Distribution of quantization levels for μ-law with 3 bits.

A compromise is the use of nonlinear relations such as the μ-law,[5,6] popular in the United States, and the A-law,[7,8,9] popular in Europe. As both are very similar, only the μ-law is described. The compression function is shown in Eq.(II-3) and its effect on the distribution of quantization levels for 3-bit PCM is shown in the graph of Fig. II-4:

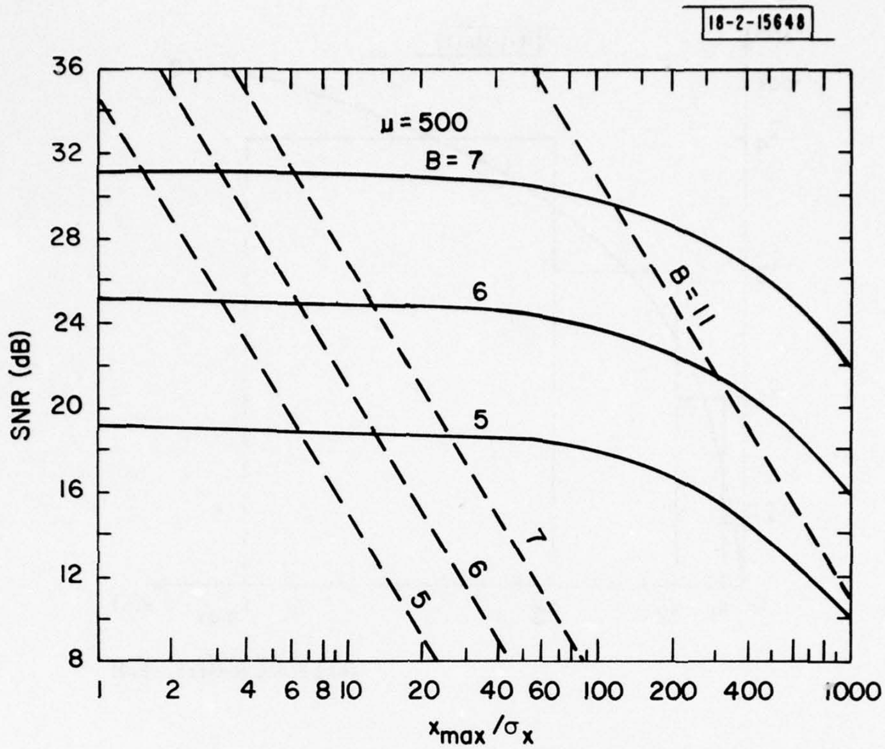
$$F[x(n)] = x_{max} \frac{\log\left[1 + \mu \frac{|x(n)|}{x_{max}}\right]}{\log[1 + \mu]} \operatorname{sgn}[x(n)] \qquad \text{(II-3)}$$

where

x_{max} = Maximum magnitude input signal permitted

μ = A system parameter .

As was desired, the quantization levels are distributed in a logarithmic fashion. As the variable μ is increased from zero – the no-compression setting – the amount of compression increases at the expense of a loss in the SNR for large amplitude inputs. A comparison of SNR for PCM and μ-law nonlinear PCM is shown in Fig. II-5 as a function of the energy in the input signal. Note that for 7 bits, the companded system maintains a SNR of 30 dB or greater until the input level is 40 dB below clipping. For use in telephone quality systems where 30-dB SNR is the standard, the 7-bit μ-law compander is comparable with an 11-bit PCM system.



(RABINER & SCHAFER, 1978)

Fig. II-5. SNR for μ-law and uniform quantization as a function of input signal level and number of quantization bits.

C. Adaptive Quantization

Instantaneous companding is one solution to the problem of encoding signals that vary in amplitude. This method is a compromise, sacrificing SNR at high input levels by spacing the quantization levels in a logarithmic manner to get a higher SNR at lower input levels. Speech, music, and most natural sounds, vary slowly in amplitude relative to the sampling rate. Adaptive quantization takes advantage of the slow variation by changing the spacing of the quantization levels as a function of the power in the input averaged over a short period of time. This modification is syllabic companding, implying adaptation over intervals comparable to syllable lengths in speech. In practice, the variation may be quite a bit quicker than the rate of speech syllables. Adaptive quantization maintains the SNR at a lower bit rate than instantaneous companding because information specifying the signal power is not coded into each sample. Rather, it is spread over the syllabic time interval related to the rate of variation of the input power.

1. Syllabic Companding

As can be done in instantaneous companding, the time-varying adaptation is often realized as preprocessing and postprocessing to a uniform quantization PCM, referred to as adaptive PCM (APCM). This scheme is depicted in Fig. II-6. The gain is varied so as to keep constant the level of the input to the quantizer.

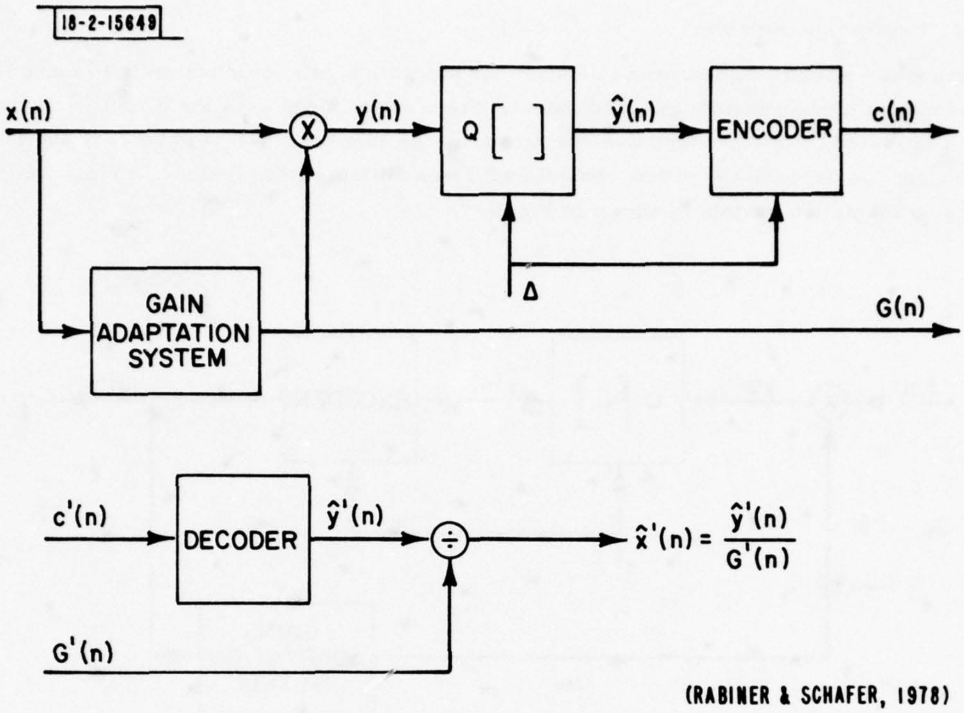


Fig. II-6. Feedforward APCM with time-varying gain.

The rate of adaptation, the rate at which the gain in Fig. II-6 is allowed to change, is an important variable. By adapting very quickly, the RMS error will be small because the input is always utilizing the full range of the PCM quantizer. The amount of information necessary to encode the gain variations, though, has increased proportionally to the frequency bandwidth of the adaptation. By adapting the gain more slowly, the bit rate for the encoding of the gain decreases. These bits may be shifted to the quantizer, compensating for not using the entire quantization range during periods that the input varies too rapidly. If the adaptation is too slow, the problem when no adaptation is used returns. In summary, it is desired to vary the gain as slowly as possible without allowing an audible lowering of the SNR and without overloading the quantizer.

Overload can be eliminated by a block adaptation scheme, a method for implementing a noncausal adaptation gain function. By filling a buffer with a block of samples, the value of the prequantization gain can be determined as a function of the samples in the buffer. Then, the gain can adapt for an attack transient or other quick increase in signal energy to avoid any overload. In this scheme, however, the error magnitude is determined by the maximum magnitude sample value in a block. During rapid and abrupt changes in input signal level, the error may be relatively large compared to the low energy portion of the input signal block, especially when the transient occurs near the beginning or end of the block. Block lengths are usually chosen in the range of 5 to 50 ms as a compromise. The constraints that are placed on the block length are discussed and quantified further in Sections III-C and IV-D in terms of the psychophysics of the auditory system.

2. Feedback Adaptation

A method for eliminating the need to encode the adaptation gain information is to make the gain be a function of the previous encoded output. Since the encoder uses the output in its processing, it is called feedback-adaptation quantization. As this information is already available to the decoder, the decoder can derive the gain with no additional information. A block diagram of a feedback-adaptation system is shown in Fig. II-7.

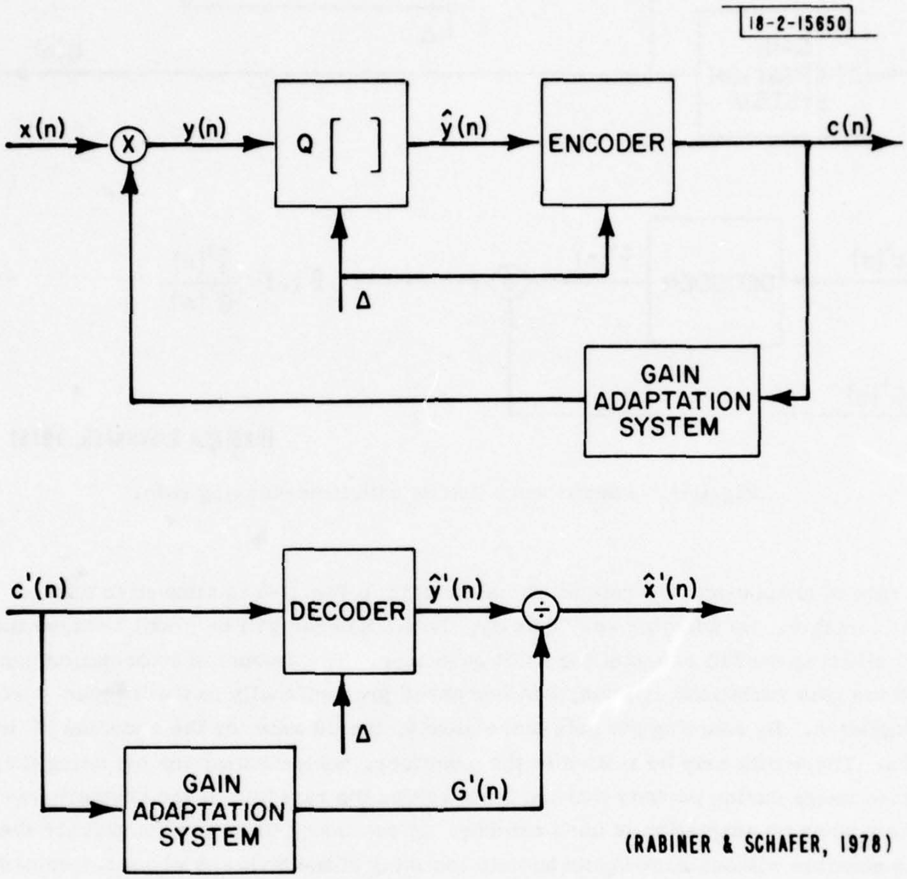


Fig. II-7. Feedback APCM with time-varying gain.

The gain can only be a function of the previous outputs because the present output is not available until after the gain has been set and the sample quantized. If the gain were permitted to change again based on the present output; i.e., as an iterative procedure, the information of the iteration would not be known by the decoder.

Signals in nature such as speech and music from instruments may have sharp attack transients where the signal energy greatly increases in less than 1 ms. In general, though, the decay of these signals is much slower. In a feedback-adaptation quantizer, large errors can be present during overload conditions. A sudden 10-dB jump in input level can produce an error from overload which is greater than the signal. It is, therefore, important to adapt very quickly by lowering the adaptation gain when the input level increases. In contrast, failure to adapt quickly to decreases in signal energy results only in a temporary reduction in the SNR by the amount of the decrease. Also, for signals with predominantly low-frequency energy, peaks in the amplitude of steady-state periods may occur as little as every 20 ms. If the gain is increased on a time scale faster than this, it must also be decreased every amplitude-peak sample. This may produce an audible pulsing of the quantization error that is more annoying than a steady-state noise would be.

D. Predictive Quantization

Predictive quantization takes advantage of the correlation that may exist between input samples. A prediction of the sample value is made and the prediction residue, the difference of the actual and the predicted signal, is quantized. If there is correlation between input samples, the prediction residue will be a signal with less power than the input. The quantizer can be adjusted for the smaller difference signal and will produce a smaller quantization error. A block diagram of a differential quantization system is shown in Fig. II-8. From the decoder we see that the output of the system with no channel errors equals the quantized difference signal plus the predicted signal as shown in Eq. (II-4):

$$\begin{aligned} x &= \tilde{x} + \hat{d} \\ &= \tilde{x} + (d + e) \\ &= \tilde{x} + [(x - \tilde{x}) + e] \\ &= x + e \quad . \end{aligned} \tag{II-4}$$

Thus, the system error, the difference between the input and the output, is exactly the error in quantizing the difference signal. Since the signal that is quantized is the prediction residue, the SNR of the output exceeds the expected SNR due to the quantizer by the amount of the prediction gain, the ratio of the input to the residual energies,[10] i.e.,

$$\begin{aligned} \text{SNR} &= \frac{P_x}{P_e} \\ &= \frac{P_x}{P_d} \frac{P_d}{P_e} \end{aligned} \tag{II-5}$$

where

P_x = Input signal power

P_d = Difference signal power

P_e = Error signal power .

Systems of the general form of Fig. II-8 are referred to as differential PCM (DPCM) systems.

1. Differential Pulse Code Modulation

For encoding speech, the improvement in SNR by using a DPCM system rather than PCM is about 6 to 8 dB with 11 dB possible for an optimized system.[11] Unfortunately, there has been little work with predictive quantization systems for music encoding so that it is difficult to quantify the SNR advantage with music input.

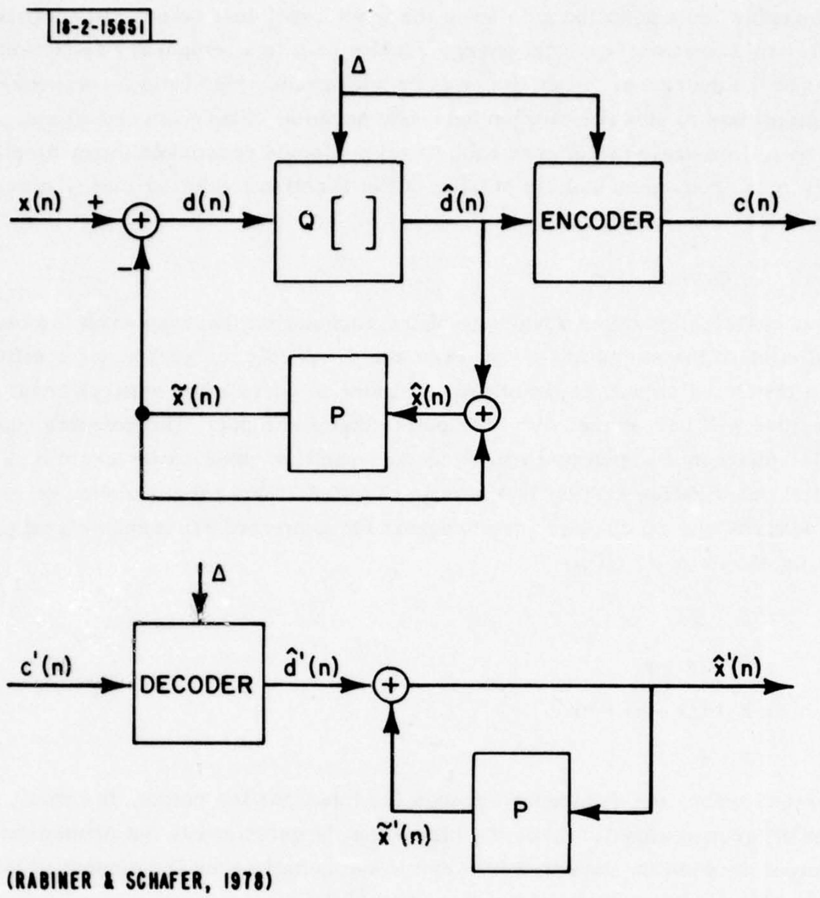


Fig. II-8. Predictive quantization system.

It is possible to combine previously described schemes with predictive quantization to achieve a combination of the improvements of each. By using an instantaneous compander such as μ-law quantization in the DPCM system, the advantages of the logarithmic distribution of quantization levels with an additional 6-dB improvement in SNR are achieved.

In adaptive differential PCM (ADPCM), the quantizer and/or predictor are permitted to adapt to the statistics of the signal. As each variable in the DPCM block diagram scales with input level, the dynamic range can be increased by varying the step size as a function of the energy in the difference signal as in APCM. By using an adaptive-quantization ADPCM system to encode speech, there is approximately a 1.5-bit (9-dB) improvement in SNR over μ-law PCM. Subjective quality as measured by listener preference shows a 2.5-bit improvement; e.g., 4-bit ADPCM is rated between 6- and 7-bit μ-law PCM in quality.

When the predictor is permitted to adapt also, the predictor gain in Eq. (II-5) can be increased additionally. Using an optimum 12th-order predictor, prediction gains of 13 dB for voiced speech and 6 dB for unvoiced speech are typical. If the input speech is pre-emphasized the sample-to-sample correlation is decreased for voiced speech sounds in exchange for the shaping of the noise spectrum by the de-emphasis filter at the processor output. With pre-emphasis, 8-dB prediction gains for both voiced and unvoiced segments are typical.[12] As a summarizing example, ADPCM with a 4th-order adaptive predictor and 1-bit adaptive quantizer can produce intelligible speech at 16 kbps comparable to 5-bit log PCM at 8-kHz sampling, a 60-percent savings.[10]

2. Delta Modulation

A simple DPCM system is delta modulation (DM), where a 1-bit quantizer and a fixed 1st-order predictor are used. By sampling the input at a rate much higher than the Nyquist rate, the input is assured to change slowly from sample to sample. The 1-bit quantizer can be set as a compromise to a level where the quantization error is small, but the differential signal is never much larger than the quantization step size. Sampling rates over 200 kbps are necessary for high-quality speech.[13]

By adapting the quantization step size to the signal, the bit rate of DM can be reduced dramatically. Adaptive delta modulation (ADM) uses a feedback adaptation scheme based on the past quantized outputs to permit tracking of the input signal at lower rates. Speech quality equal to 4-kHz bandwidth, 7-bit log PCM can be produced at the same bit rate using ADM, 56 kbps.

E. Time-Frequency Domain Quantization

The object of all schemes for the encoding of speech and audio is to approximate the signal with the smallest error for a given information rate. For speech, models of the speech production process permit tailoring of the system and adaptation of the parameters for increased SNR. For instrumental music and other natural sounds, signal statistics are available but are not as predictable and, hence, do not permit as much improvement as with speech inputs.

The level of the error of an encoding system is usually measured in terms of SNR with various inputs. Most systems are optimized to achieve the highest SNR over the range of allowable input signals. After the system is optimized, it is tested for audibility of the encoding error, quality, and intelligibility. By initially designing the encoder to minimize the audibility of the error, further improvements may be made.

The systems just described are waveform coders, adapting and predicting on a time-sample to time-sample basis for the full-bandwidth time waveform. The encoding error is a noise process with time-varying variance and a flat-power spectral density that may be postfiltered by de-emphasis. The audibility of the error, however, is dependent on the dynamic temporal and spectral relation of the signal and error. Time-frequency domain quantization systems attempt to transform the signal into a domain where quantization is better matched to audition. Although the psychophysics of the auditory system are detailed in Chapter 3, a brief review of existing time-frequency domain encoders is given in this Section.

1. Sub-Band Coding

The sub-band coder divides the signal into several bandpass frequency channels, typically 4 to 8 in number, each to be quantized separately. Use of APCM quantization for each band maintains a constant SNR over a large range of input levels. A block diagram of the sub-band coding system is shown in Fig. II-9.

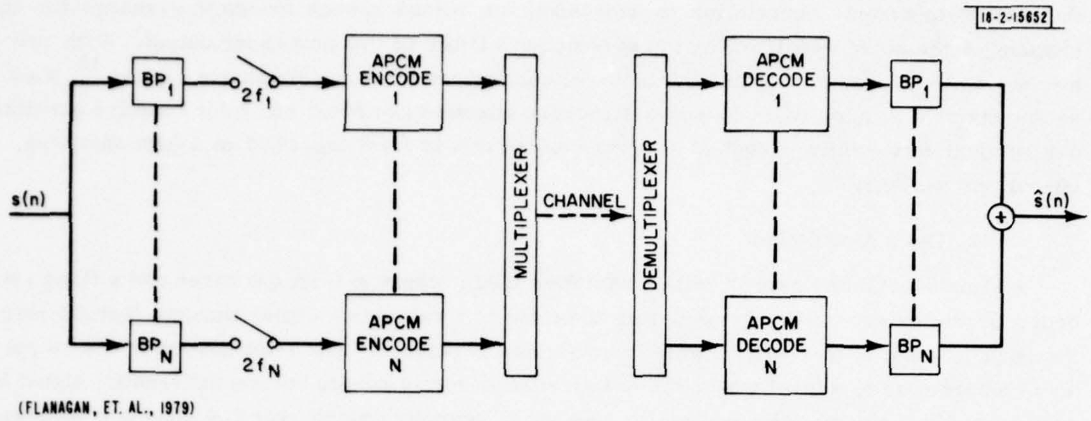


Fig. II-9. Sub-band coder.

By coding each sub-band independently, the quantization error from a band will be constrained to be in that band. When a signal with predominantly low frequency energy is coded by the waveform coders of the previous sections, the quantization error is white noise. The high-frequency components of the noise will then be audible unless the SNR is much greater than 60 dB.[14] The quantization error of the sub-band coder is restricted to frequencies close to the signal-frequency components. High-frequency channels with low-signal energy will contribute only small errors with energy proportional to the signal energy in that band. An SNR of 40 dB may be sufficient to render the noise inaudible.

Another advantage of the sub-band system is that the bit rate and, hence, the SNR of each band may be chosen independently of the other bands. In speech signals, significant high-frequency energy is present only for unvoiced speech sounds. Since this is a noise-like signal, greater quantization error is tolerable, especially if it is shaped to the speech spectrum. Thus, fewer bits can be used to encode the upper-frequency channels.

For speech coding at 16 kbps, sub-band coding has a slightly higher SNR – 11.2 dB – than ADPCM – 10.9 dB. Subjectively, however, it is comparable to 22-kbps ADPCM.[15] It is unclear how sub-band coding compares with the optimized adaptive quantizer and adaptive predictor ADPCM at 16 kbps that has an SNR of 17 dB.[5] Note, however, that this ADPCM receives significant predictor gain from optimum prediction of the speech waveform. If signals other than speech were used as input, the predictor gain would decrease greatly while the sub-band coder would not be degraded significantly.

2. Transform Coding

Transform coding is another technique to match the quantization to the short-time Fourier analysis that is performed by the auditory system. Whereas the sub-band coding system quantizes time samples from a window in frequency (the bandpass filter output), transform coding quantizes frequency samples from a window in time. Although the effects are similar, the implementations differ due to the characteristics of the specific frequency transformation used in the transform coding and differing adaptation strategies due to the statistics of the signals to be quantized.

Adaptive transform coding (ATC), transform coding with an adaptive quantization-bit distribution strategy, yields a 3- to 6-dB advantage over ADPCM when optimized for speech.[16] Perceptually, however, it is quite different. The degradations introduced by the encoding are not perceived as noise. They are manifested as changes in the quality of the signal. ATC can produce toll-quality (telephone quality) speech at a bit rate of 16 kbps for 3.2-kHz bandwidth speech.[17]

III. PERCEPTUAL REQUIREMENTS OF A DIGITAL ENCODER

A. Introduction

The performance of any audio system can only be judged by how it sounds. On all but the highest quality sound reproduction system, music through 16-bit PCM encoding systems will sound identical to the original signal. Experiments by the British Broadcasting Corporation (BBC) and German Post Office show that 13-bit PCM encoding yields acceptable quality for many applications in radio broadcast.[7,18,19,20] If, however, the digital signal is to be processed further, the sum of the errors due to 13-bit encoding and subsequent processing may sound degraded with respect to the original and 16-bit PCM signals.

To decide when errors will be audible and, therefore, what is important in the design of an encoding system, it is necessary to have an understanding of the capabilities and limitations of the auditory system. Although there are no complete models of the auditory system, there is a large body of literature detailing experiments, relating various psychoacoustic phenomena, and modeling certain aspects of the auditory system. This literature, along with some simple experiments, can be used to specify the perceptual requirements of a digital encoder for audio signals.

In this Chapter, the concepts of masking and critical bands are reviewed briefly and quantified with reference to the results of experiments from the literature. It will be hypothesized that by tailoring the spectral and temporal aspects of the error signal from a digital encoder, it is possible to specify conditions under which the error is rendered inaudible when presented along with the audio signal. Thus, the output of the encoder will sound identical to the input.

B. Masking and Critical Bands

By its very nature, the approximation of a continuous audio waveform by a discrete digital bit stream will have an error. Depending on the encoding system, this error or noise signal may or may not be correlated with the audio signal. The quality of the system is dependent on the audibility of the encoding noise.

The approximate range of normal human hearing spans from 20 Hz to 20 kHz in frequency. As seen in Fig. III-1, the threshold of audibility for pure tones varies from a minimum at 3.5 kHz of −4-dB sound pressure level (SPL) relative to the standard pressure of 0.0002 dyne/sq cm, the normal threshold at 1 kHz.[21] The threshold rises to a maximum of 73-dB SPL at 20 Hz. Levels above 140-dB SPL are usually painful, representing the practical upper amplitude limit of hearing. To try to design for this dynamic range of 144 dB would require major advances in the state-of-the-art in electronic instrumentation. Even if this range is restricted to 100 dB, a more practical limit with respect to most speech and audio signals, a linear PCM encoder would require 16-bit coding. Several digital encoding systems, however, do use 16-bit PCM to achieve large dynamic range and very low noise.

The thresholds shown in Fig. III-1 are for pure tones in isolation. When a tone is presented along with other signals, the threshold may be raised in a manner dependent on the spectral and temporal characteristics of the other signal. This is masking, the phenomenon of one audio signal inhibiting the detection of another audio signal. Figure III-2 shows how a complex signal composed of two tones is preceived. The graph is for a 1200-Hz, 80-dB SPL primary tone and is plotted as a function of secondary-tone frequency and level.[22,23] Under these conditions, the masked threshold of the secondary tone – the minimum level where the secondary or target tone

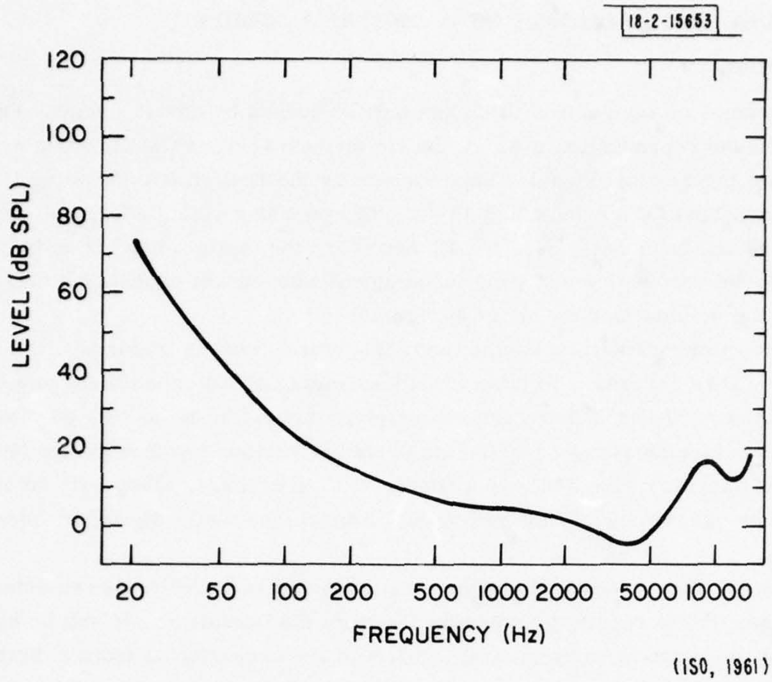


Fig. III-1. Threshold of audibility or pure tones.

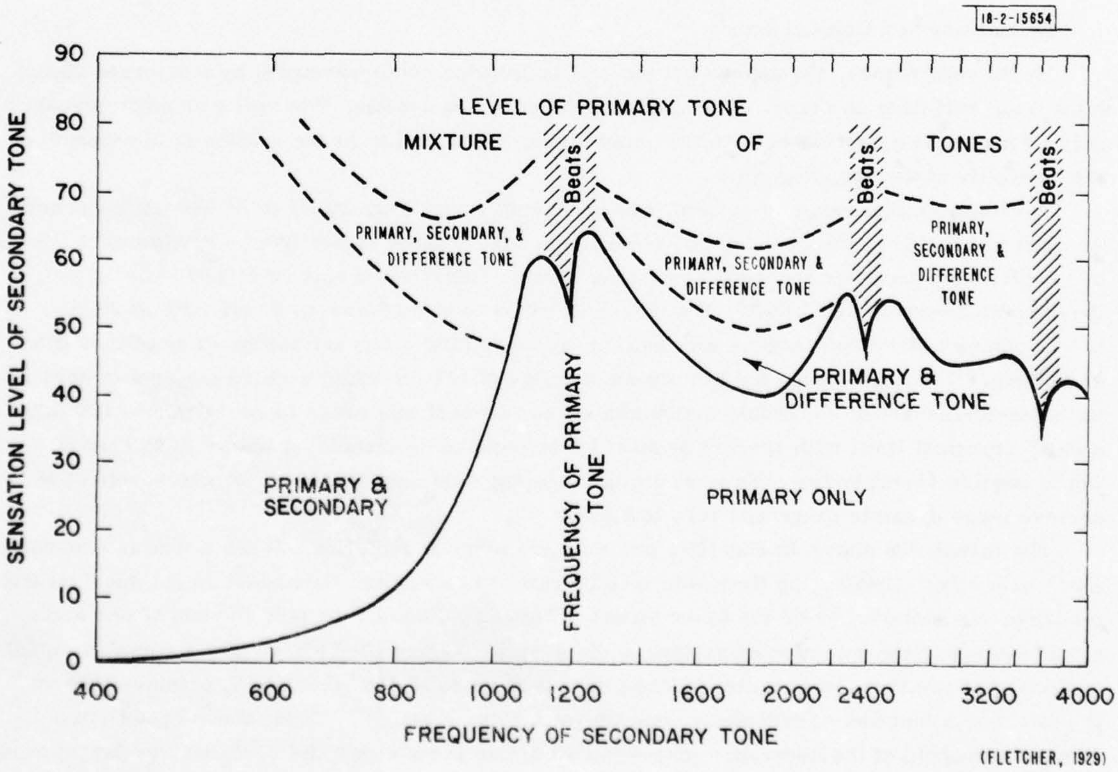


Fig. III-2. Perception of a two-tone signal.

is perceived along with the primary or masking tone – is higher than the unmasked threshold from Fig. III-1. The amount that the threshold is raised is the amount of masking due to the masking tone. Note that the amount of masking is greater for frequencies near the masking frequency and that there is little masking at frequencies below the masking frequency.

By replacing the masking tone by a narrow band of noise, the masked threshold is no longer dependent on the exact frequency relation of masker and target. Figure III-3 shows the masking pattern of a 90-Hz narrowband noise signal presented at several sound pressure levels.[24,25] The notches due to "beat" phenomena in Fig. III-2 have been replaced by peaks in the masking audiogram. The general shape of the curves indicates that the amount of masking in the vicinity of the center frequency of the band of noise is approximately linearly related to the noise power, i.e., a 10-dB increase in noise power results in a 10-dB increase in the sinusoid masked threshold. For the experiment in Fig. III-3, the masked threshold of a 410-Hz tone centered in a 70-dB SPL noise masker is the unmasked threshold (7-dB SPL) plus the amount of masking (53 dB) for a total of a 60-dB, SPL-masked threshold. Thus, the level of the tone must be within 10 dB of the masker power to be audible at this frequency.

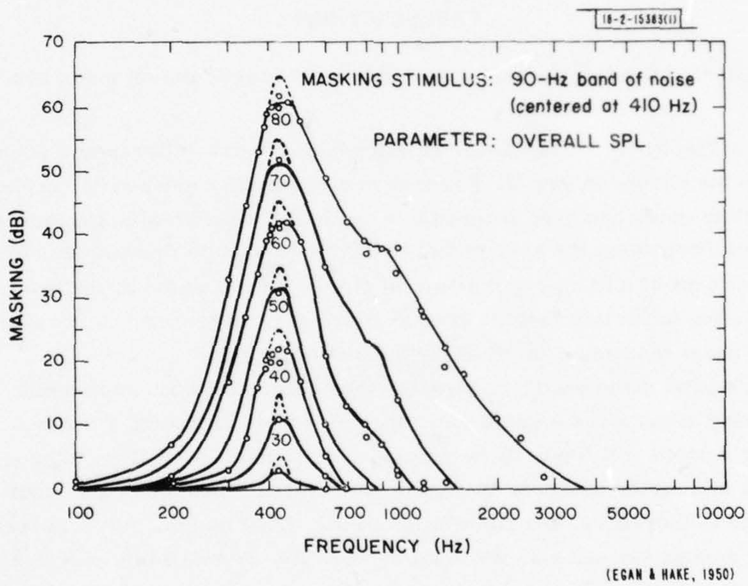


Fig. III-3. Masking audiogram of a narrow band of noise.

The frequency range of masking, as well as many other psychoacoustic phenomena such as loudness perception and phase audibility is related to the critical band, the bandwidth where there is sudden change in observed subjective responses. The critical band is often defined as the range of frequencies of a noise signal that contribute to the masking of a pure tone centered in frequency in the band of noise.[25,26,27] It is measured by masking a tone by bandpass-filtered white noise. The bandwidth of the masker is decreased until the masked threshold of the tone starts to decrease. The bandwidth of the noise masker at that point is defined to be the critical bandwidth at the frequency of the sinusoid. A graph of critical bandwidth as a function of frequency and compared to third octaves, sixth octaves, and 5-percent articulation index[13,28] is shown in Fig. III-4.

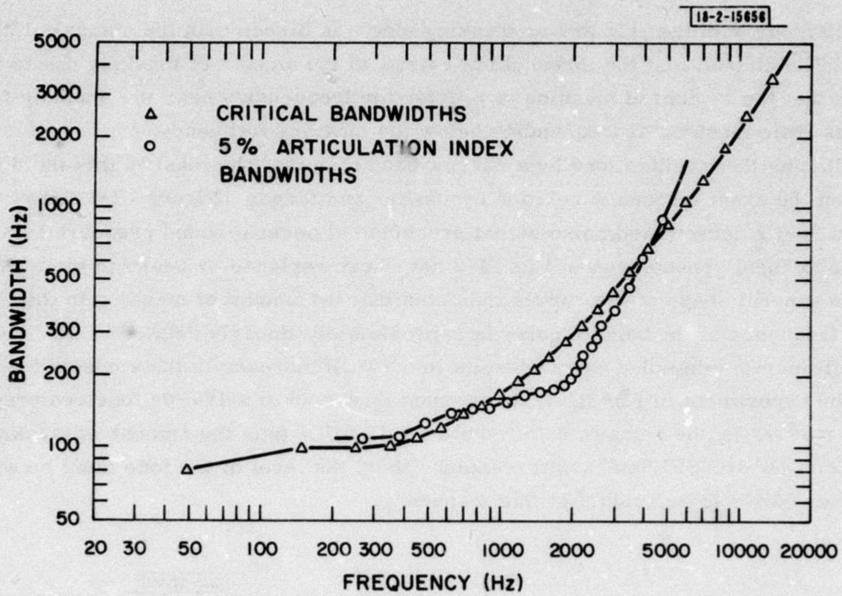


Fig. III-4. Bandwidth of critical bands and 5% articulation index bands.

It is shown in Fig. III-5 that the linear relationship of masker power and amount of masking conjectured from the curves in Fig. III-3 is true and essentially independent of the frequency of the tone.[25,29] White noise has been filtered to a critical bandwidth at six frequencies. At several levels at each frequency, the narrow band of noise is used to mask a sinusoid centered in the noise. The amount of masking of the tone is plotted as a function of the level of the critical band of noise relative to its threshold. At each frequency, an increase in masker level is accompanied by an equal increment in dB of the amount of masking.

For use in a digital audio encoder, signals other than narrowband noise must be considered as possible masking signals and signals other than sinusoids as masking targets. The error from an encoding system will typically be a noise-like signal. A PCM encoder is a time-invariant nonlinear system that yields an error that is a deterministic function of the input. As the number of bits per sample is increased, the correlation of the signal and the error is reduced. Perceptually, the error sounds the same as white noise when five or more bits per sample are used. Thus, it is important to consider the masking of noise signals. Consistent with the masking curves in Fig. III-3, there is very little masking of wideband white noise by narrowband signals such as tones.[14,30,31] Narrowband signals would, however, be expected to mask narrowband noise.

Not finding any pertinent experiments in the literature, the following experiment was performed. For each of 18 frequencies, narrowband noise was masked by a sinusoid at a listening level of approximately 70-dB SPL at the center frequency of the band of noise. The noise was obtained by passing wideband noise through a 4-pole Butterworth filter. The bandwidths of the filters, shown in Table III-1, are approximately of critical bandwidth. The sinusoid and noise signals were presented together monaurally over headphones. Each subject varied the amplitude of the noise until it was at the minimum audible level. The results of three subjects were

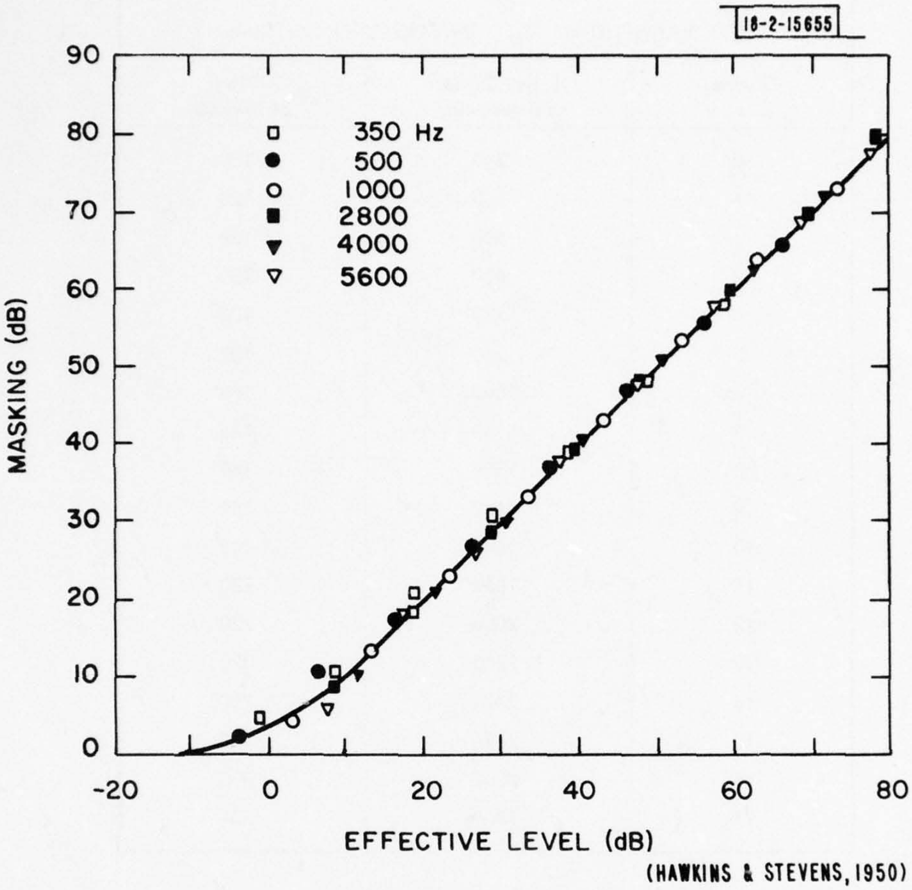


Fig. III-5. Relation between the masking by white noise and the effective level of the noise. The effective level is the power in the critical band around the masked sinusoid.

TABLE III-1

FILTER BANDWIDTHS USED IN MASKING EXPERIMENT

Channel Number	Filter Center Frequencies	Filter Bandwidths
0	240	130
1	360	130
2	480	130
3	600	130
4	720	130
5	840	130
6	1000	165
7	1150	165
8	1300	165
9	1450	165
10	1600	165
11	1800	220
12	2000	220
13	2200	220
14	2400	220
15	2700	330
16	3000	330
17	3300	330

averaged and are presented in Fig. III-6. The SNR necessary for the noise to be masked varies with frequency, reaching a maximum of 28 dB at 1150 Hz. Thus, the frequency at which the least amount of masking is present is 1150 Hz. The maximum variation of a subject from the average was 4 dB. This indicates that tones are effective at masking narrowband noise, but the amount of masking is less than the amount of masking for narrowband noise masking of tones.

C. Nonsimultaneous Masking

The previous experiments all use signals that are presented simultaneously. Judgments of audibility were made when the signals were in the steady state and not during transients. Masking, however, also occurs when the stimuli are not presented simultaneously. A masking signal can mask sounds occurring before, referred to as backward or premasking, or after, referred to as forward or postmasking. Although the time period for nonsimultaneous masking is short, less than 100 ms, it is very important to the question of digital encoding. By not requiring the system to adapt instantaneously to transients in the input waveform, a large saving in the amount of information to specify the signal and, hence, a lowering of the bit rate can be obtained. This principle is used in the design of compressors, limiters, and automatic gain controls for audio recording and broadcast industries.

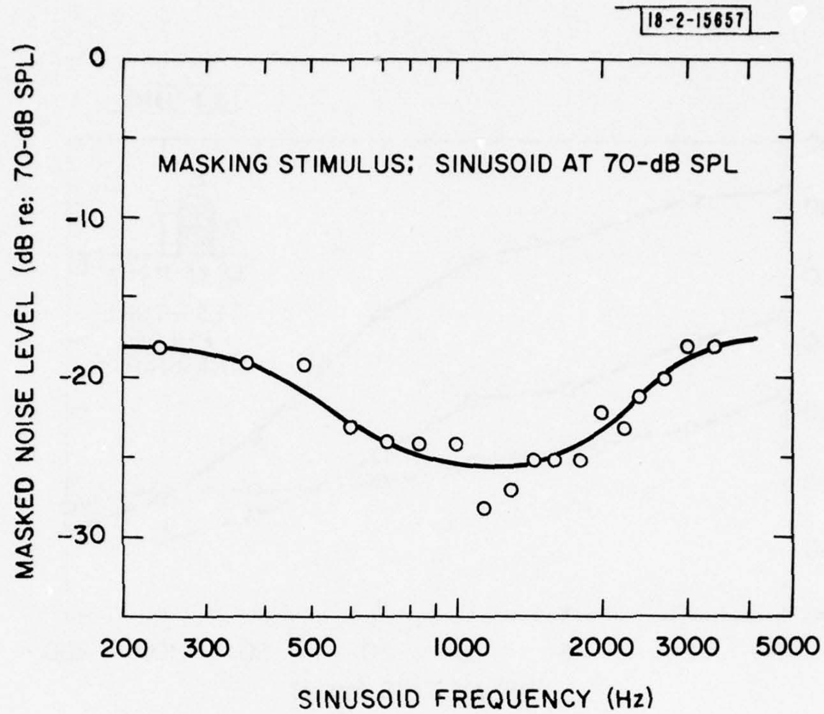


Fig. III-6. Ratio of sinusoid masker level to critical bandwidth noise level at masked threshold.

In Fig. III-7 the masking of short tone bursts by a single, critical-bandwidth, noise masker is shown as a function of the timing of the tone burst in relation to the masker.[32] The masker is narrowband noise centered at 8.5 kHz, 1.8 kHz in bandwidth, and presented at 70-dB SPL for a duration of 500 ms. The tone bursts at frequencies of 6.5, 8.5, and 11 kHz, were 1 ms in duration. The unmasked thresholds of the tone bursts are also shown for reference. For 8.5-kHz tone bursts centered in frequency in the band of noise, the nonsimultaneous masked threshold is within 20 dB of the simultaneous masked threshold when the tone burst occurs within 10 ms before or 30 ms after the noise-masker burst.

If the duration of the tone bursts are increased slightly the shape of the curves remain unchanged but are shifted downward; i.e., there is less masking.[32] This shift is by approximately the same factor as the increase in energy of the burst in agreement with the short-time temporal integration in the auditory system.[13] Thus, increasing the duration from 1 to 10 ms lowers the masked threshold by about 10 dB; i.e., to 15 dB below the masker level.

The results of these experiments are summarized in Fig. III-8. The backward, forward, and simultaneous masking curves have been combined to reflect the total transient masking pattern of the tone burst by the critical-band noise masker.

D. Summary

The masking data indicate that a sinusoid raises the masked threshold of a narrowband noise to within 28 dB of the sinusoid level. Other experiments imply that nonsinusoid maskers having a greater spectral spread, provide even more masking.

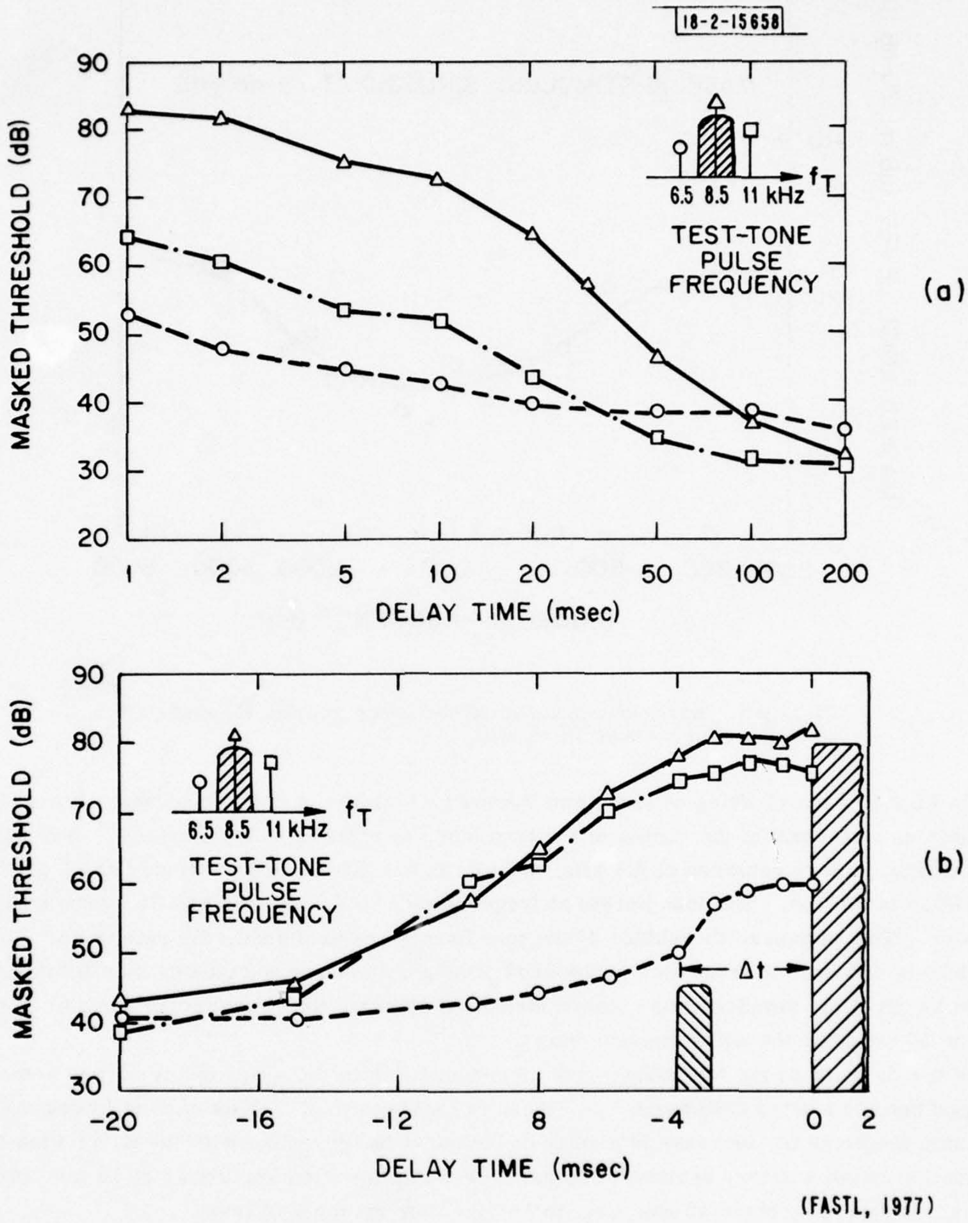


Fig. III-7. Nonsimultaneous masking of tone pulses by critical bandwidth noise masker bursts: (a) forward masking, (b) backward masking.

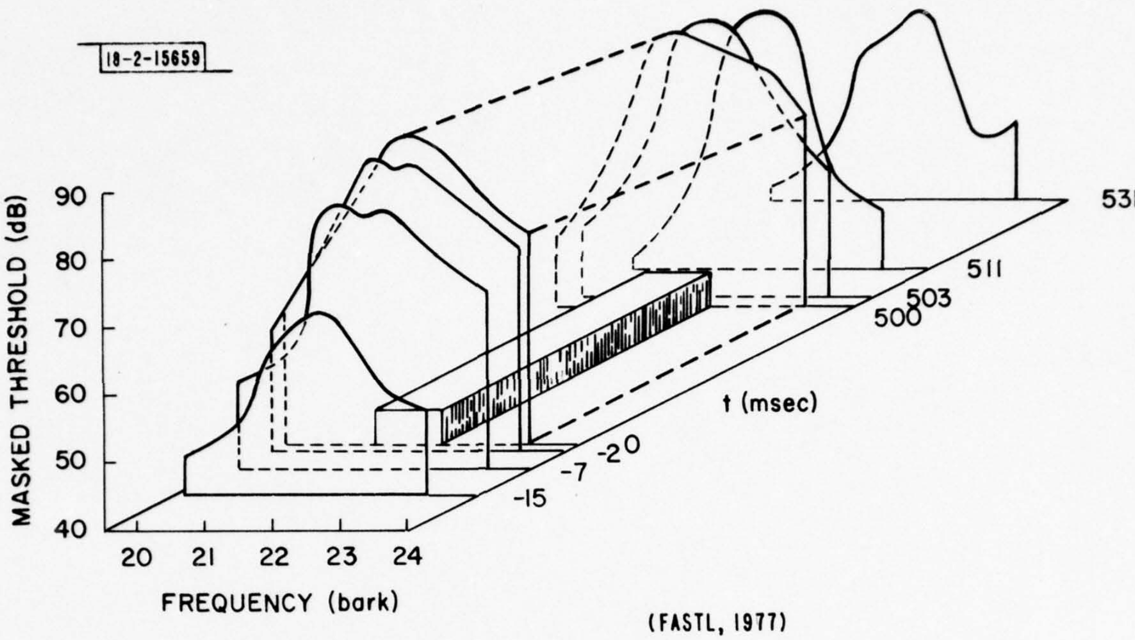


Fig. III-8. Transient masking pattern of a single, critical bandwidth, noise-masker impulse.

The signals that are to be encoded – speech, music, and other audio signals – are not narrowband signals as are the tone and noise maskers of the previously described experiments. It is possible to divide the wideband audio signal into several narrowband signals by filtering techniques. It is hypothesized that if the perceptual requirements for the encoding error to be masked are met for every narrowband frequency section of the output signal, then the error will still be masked when all of the frequency bands are presented together. This hypothesis simply states that masking is additive for masker-target pairs that are nonoverlapping in frequency.

To meet these requirements efficiently, a system should have an error that adapts to the input signal. The noise spectrum, as measured by a short-time Fourier transform or through a bank of bandpass filters, should be shaped so that the SNR in every frequency band of critical bandwidth is adequate for masking. Temporally, the system should adapt quickly enough such that errors occurring shortly before or after transients will also be masked.

IV. DESIGN OF THE DIGITAL ENCODER

A. Introduction

By shaping dynamically the spectrum of the encoder error, it is possible to render the error inaudible by the masking action of the input signal. The masking curves in Chapter III show that when the noise is restricted to a narrow band a sinusoid will mask the noise if the SNR is sufficient; i.e., greater than that shown in Fig. III-6, or a maximum of 28 dB for any frequency band. For nonsteady-state signals, temporal restrictions on the error signal are implied by the nonsimultaneous masking results. These experiments are used to define a time-frequency domain similar to the short-time spectral analysis domain used for analysis by the auditory system. By transforming signals into this domain, the quantization process can be better matched to audition.

Two basic techniques are currently used for the signal transformation, as was noted in Section III-E. The subband coder uses a bank of bandpass filters to separate the signal into several independent frequency channels for quantization. The transform coder uses a discrete cosine transform to form a short-time spectral analysis of a windowed time segment. Although either method performs the desired transformation, bandpass filtering was chosen since the temporal and spectral parameters in the implementation are related closely to auditory performance parameters. In this Chapter, the block diagram of the system is presented and the implementation of the blocks described. Relation of the parameters to the perceptual requirements analyzed in Chapter III are discussed. Details of the signal processing aspects of the implementation are discussed in Appendix I.

B. Block Diagram of the Encoding System

The structure of the encoding system is shown in Fig. IV-1. The audio signal is filtered into a set of 24 contiguous frequency bands that cover the audible frequency range. For input signals that do not require system response to 15 kHz, fewer bands are used. For speech inputs, 17 filter channels cover the range to 4.1 kHz. By quantizing each channel independently, the quantization error can be restricted to the frequency band of that channel. Quantization accuracy and temporal adaptation characteristics may also be chosen according to the discrimination ability of the auditory system in each frequency range. To ensure that the quantization error is masked by the signal, each channel is approximately of critical bandwidth or smaller.

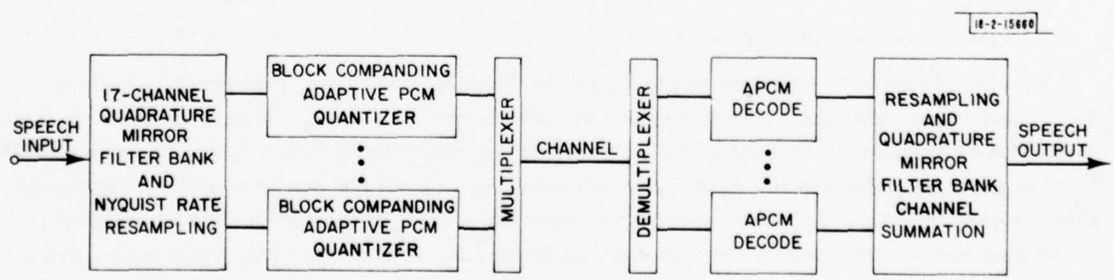


Fig. IV-1. Digital encoding system.

After the signal is filtered into narrow frequency bands, each channel may be resampled at a lower sampling rate commensurate with the channel bandwidth; i.e., at the Nyguist rate of the channel. If the channels are contiguous and nonoverlapping, the sum of the sampling rates of the channels would equal the original sampling rate. Unfortunately, an ideal bandpass filter is not a causal function and is, therefore, not realizable. Implementation considerations of the filter bank are discussed in Section IV-C.

One implementation solution that achieves the desired spectral shaping is to eliminate the resampling operations before and after quantization. Sampling would then be at a rate much higher than the Nyguist rate as occurs in a delta modulation system. Assuming that the quantization error can be approximated by a wideband white-noise process, filtering to the signal bandwidth after quantization will remove out-of-band noise. The SNR will be increased by the ratio of the noise bandwidth to the channel bandwidth. The total bit rate for this encoding will also increase by the same factor as there are more samples per second to be quantized. For each doubling of the sampling rate over the Nyquist rate and, hence, doubling of the bit rate (assuming no change in the number of quantization bits per sample), there is a 3-dB increase in the SNR resulting from filtering out half of the noise energy. Doubling of the bit rate by maintaining the sampling at the Nyquist rate and increasing the accuracy of the quantization by doubling the number of bits per sample, however, would result in the squaring of the SNR a doubling of the SNR expressed in dB. In general, therefore, unless significant predictor gains can be achieved with the over-sampled signal by using a DPCM-quantization scheme, sampling a channel at higher than its Nyguist rate does not represent an efficient allocation of the bits available for encoding.

For the quantization error to be masked by the input signal, it is necessary to shape the spectrum of the error dynamically as a function of the short-time spectrum of the signal. This implies that the quantization algorithm must maintain relative accuracy over a large dynamic range. The choice between instantaneous companding and adaptive quantization algorithms is discussed in Section IV-D.

After quantization, each channel is resampled up to original sampling rate. This resampling process causes replicas of the narrowband signal spectrum in other parts of the frequency range of the system, parts covered by other channels. Filtering each channel to its original frequency band before summing with the other channels eliminates these images. If the signal-processing requirements have been followed, the encoding error present in the reconstructed output will be due only to the quantization and not to artifacts of the implementation.

C. Design of the Filter Bank

Several issues are important in the design of the filter bank from the viewpoint both of theory and implementation. The bandwidth of each channel in the filter bank should be a critical bandwidth or smaller so that the encoding error of the channel is masked by the channel signal. For the greatest reduction in the bit rate, it is desired to have the lowest possible resampling rate for each channel. All realizable filters have finite transition bands and finite attenuation in the stop bands. Because of the aliasing that occurs for under-sampling, the sample rate for each channel must be at least twice the bandwidth to the stop-band edges. The resultant aliasing will be due only to signals at frequencies leaking through the filter stop band. By using filters with enough stop-band attenuation, this aliasing will be inaudible. Since it is necessary for the sum of the channel signals to sound the same as the original input signal, the passbands must be continguous. Thus, the narrower the transition bands of each filter, the less over-sampling will be necessary to avoid audible aliasing. Although high-order recursive (IIR) digital filters (such as 16th-order elliptic filters) have good specifications in terms of transition bandwidth for the channel bandwidth in question, they have rapid phase fluctuations at the band edges. The resultant phase distortion and ripple in the magnitude of the frequency response is often audible and, therefore, not acceptable. The use of nonrecursive (FIR) filters having linear phase can eliminate this problem. Nonrecursive filters, however, have only zeros in their z-plane response and require a much higher order than recursive filters for similar-width transition bands. A windowed bandpass-filter design for a signal with a sampling rate of 30 kHz having a 50-Hz transition band would require an FIR filter with an approximate length of 1400.

Several schemes for implementing the filter bank have recently been demonstrated.[15,33,34] A filter bank consisting of equally spaced, frequency-translated replicas of a prototype lowpass filter, leads to a simple condition for the sum of the channels to equal the input.[34] Each of the bandpass filters has an impulse response as shown in Eq. (IV-1), where h(n) is the impulse response of a lowpass filter with bilateral bandwidth equal to the bandwidth of the bandpass filter:

$$h_k(n) = h(n) \cos[\omega_k n] \quad . \tag{IV-1}$$

An implementation of the kth filter is shown in Fig. IV-2. It is now necessary to design only one lowpass filter to produce the bank of bandpass filters. Choosing the center frequencies of the N channels so that the filters are equally spaced and cover the frequency band, the impulse response of the system is Eq. (IV-2):

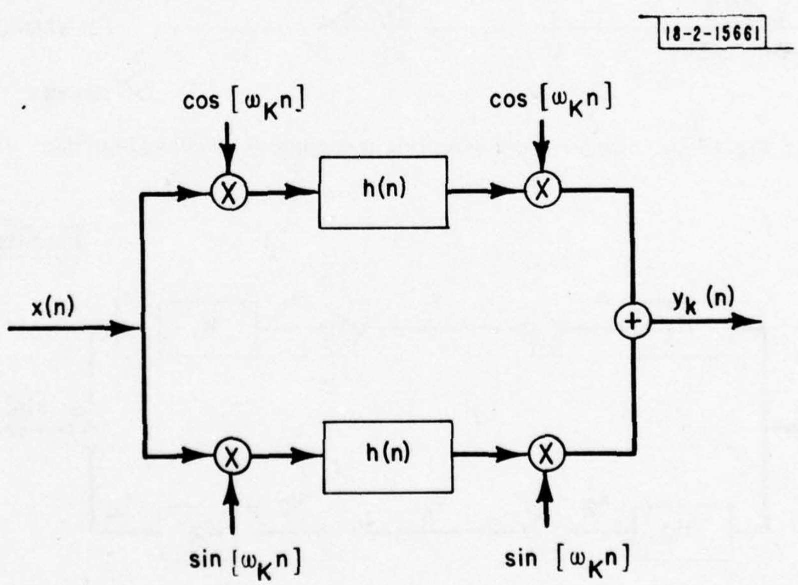


Fig. IV-2. Implementation of one channel of a filterbank.

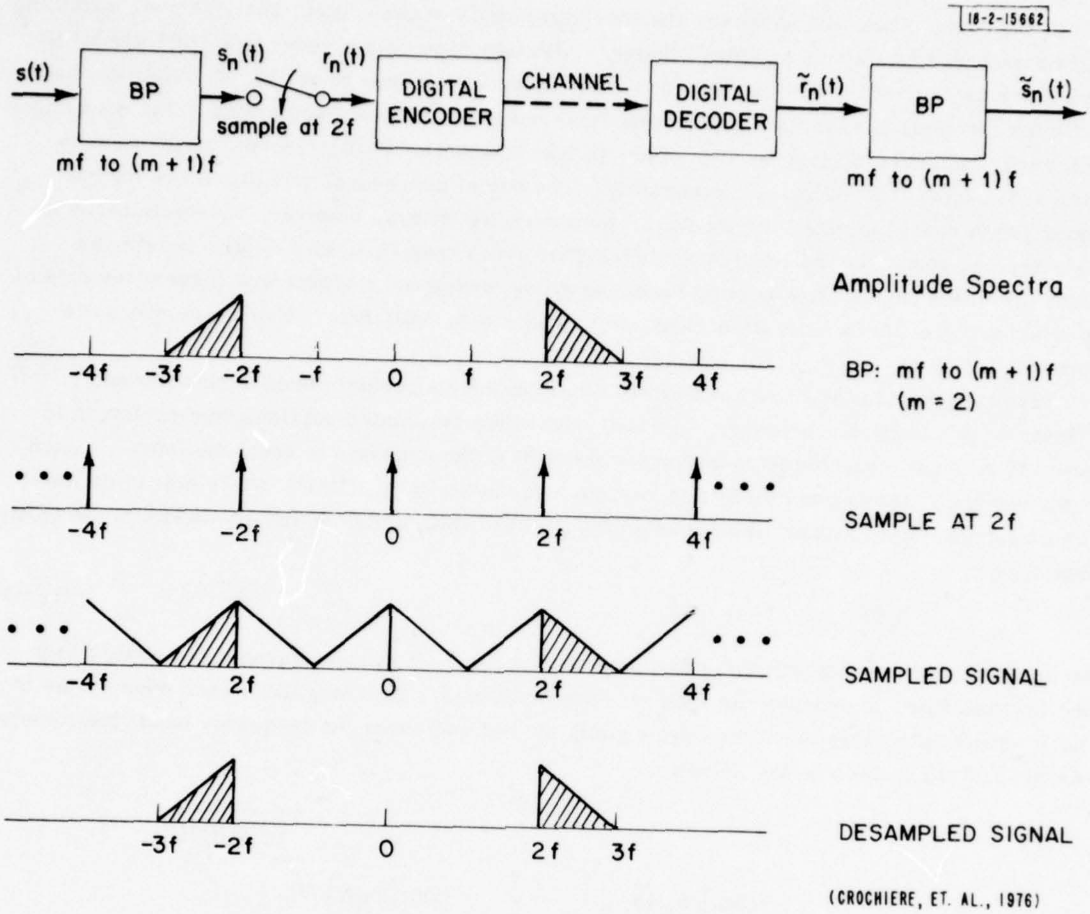


Fig. IV-3. Integer-band sampling technique for subband coding.

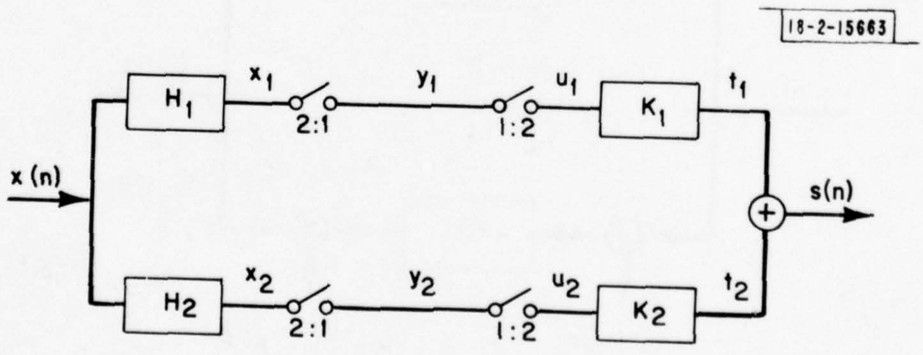


Fig. IV-4. Two-channel quadrature mirror-filter decomposition.

$$g(n) = \sum_{n=1}^{(N-1)/2} h(n) \cos [2\pi kn/N]$$

$$= h(n) \sum_{n=1}^{(N-1)/2} \cos [2\pi kn/N]$$

$$= h(n)\, d(n) \quad . \tag{IV-2}$$

Since d(n), the sum of equally frequency-spaced cosines in Eq. (IV-2), is a pulse train with a spacing of n = N, the composite system response is just a sampling of the prototype lowpass filter. The lowpass filter is designed such that when sampled, it is an impulse, the desired impulse response of the system. An efficient implementation of this system has been produced using the fast Fourier transform (FFT) algorithm.[35] It is difficult, however, to adapt this system implementation to the case of unequal bandwidth channels.

Another system that implements bandpass filters uses integer-band sampling.[15,36] A band-limited signal may be sampled at the Nyquist rate of twice its bandwidth. This is often implemented by modulating the signal so that the band of interest is centered at zero frequency. A lowpass filter is then used and the signal can be down-sampled without aliasing. The integer-band sampling scheme does not modulate the signal bands to zero frequency. Bandpass filters are used and each band is sampled at its Nyquist rate. When the band is not a lowpass signal channel, care must be taken so that the sampling does not cause aliasing of the signal back into its frequency band even though the sampling rate is sufficient. If the band limits are chosen such that the low-frequency cutoff is an integer multiple of the channel bandwidth, the signal may be down-sampled without modulation to baseband with no resultant aliasing. This restriction is a fundamental limitation with the integer-band sampling scheme. The system implementation is shown in Fig. IV-3. If the bandpass filter is a nonrecursive filter, efficiency can be gained by computing only those values that will be sampled. For a 100-Hz bandpass filter to be down-sampled from 10 kHz to 200 Hz, this implies calculating one sample of each fifty. If recursive filters are used, every sample must be computed since the filter bases the output on past outputs as well as inputs. The main advantage of the integer-band sampling system is simplicity and smaller computational loads because modulation to the zero frequency is not necessary.

A major disadvantage of these filter-bank systems is that the filters must be very sharp so that the amount of over-sampling necessary to avoid aliasing is small. A recently developed filtering technique, quadrature mirror filtering,[37,38] permits the realization of a filter bank with no aliasing and with total sampling rate equal to the sampling rate of the signal before filtering.

1. Quadrature Mirror Filtering

The basic quadrature mirror-filter technique is designed to divide the digital frequency spectrum into two equal parts. Each band is sampled at half the rate of the original signal for quantization, coding, and transmission. The signals are then resampled back up to the original rate and filtered again before summing. Since the filters are not ideal filters, there is some aliasing after the down-sampling. Because of the special relationship of the filters, when the two bands are summed, the components due to the aliasing of one band cancel the components due to aliasing of the other band. Thus, there is no aliasing in the resultant signal. The structure of the basic filter block is shown in Fig. IV-4. A summary of the important aspects of quadrature mirror filtering is given in this Chapter. A more detailed discussion can be found in Appendix I.

When the four filters shown in Fig. IV-4 are defined by their relationship to the prototype lowpass filter, h(n), as in Eq. (IV-3), the aliasing will vanish:

$$h_1(n) = h(n) \tag{IV-3a}$$

$$h_2(n) = (-1)^n\, h(n) \tag{IV-3b}$$

$$k_1(n) = h(n) \tag{IV-3c}$$

$$k_2(n) = -(-1)^n\, h(n) \quad . \tag{IV-3d}$$

These relations in terms of the z-transforms of the filters can be written as:

$$H_1(z) = H(z) \tag{IV-4a}$$

$$H_2(z) = H(-z) \tag{IV-4b}$$

$$K_1(z) = H(z) \tag{IV-4c}$$

$$K_2(z) = -H(-z) \quad . \tag{IV-4d}$$

The replacement of the parameter z by $-z$ represents a rotation by an angle of π in the z-plane. This is a shift of π in the Fourier transform of these signals. Thus, $H(-z)$ is a highpass filter with a frequency response being a shifted replica of the frequency response of $H(z)$, a lowpass filter. If h(n) is a real function, the magnitude of its frequency response is an even function. A shift of π will be equivalent to a reflection of the magnitude about the point, $w = \pi/2$. Hence, $H(-z)$ is the mirror filter of $H(z)$.

The output of the system, S(z), shows that the aliasing components have cancelled, leaving only a linear filtered term as desired:

$$S(z) = \frac{1}{2}\,[H^2(z) - H^2(-z)]\,X(z) \quad . \tag{IV-5}$$

Note that there have been no restrictions up to this point on the prototype filter, h(n), only on the relationships of the other filters to h(n). Design of this filter is discussed in Section IV-C-2. If h(n) is a good approximation to the ideal half-band lowpass filter, the signal will be divided into two equal bandwidth frequency bands by it and its highpass mirror filter.

In summary, if the relations of Eqs. (IV-3) and (IV-4) are used, the signal components due to aliasing are cancelled, leaving only linearly filtered signal components in the output. If the filter is chosen such that:

$$\left|\frac{1}{2}\,[H^2(z) - H^2(-z)]\right| = 1 \tag{IV-6}$$

and the phase is linear, then the system will be an identity system except for a delay.

The quadrature mirror-filter system can be simply extended to divide the frequency range into more than two bands. A system that implements a decomposition into four bands is shown in Fig. IV-5. If the conditions on the filters are met such that the basic two-band system is an identity system, then introduction of that identity system in the middle of another system will

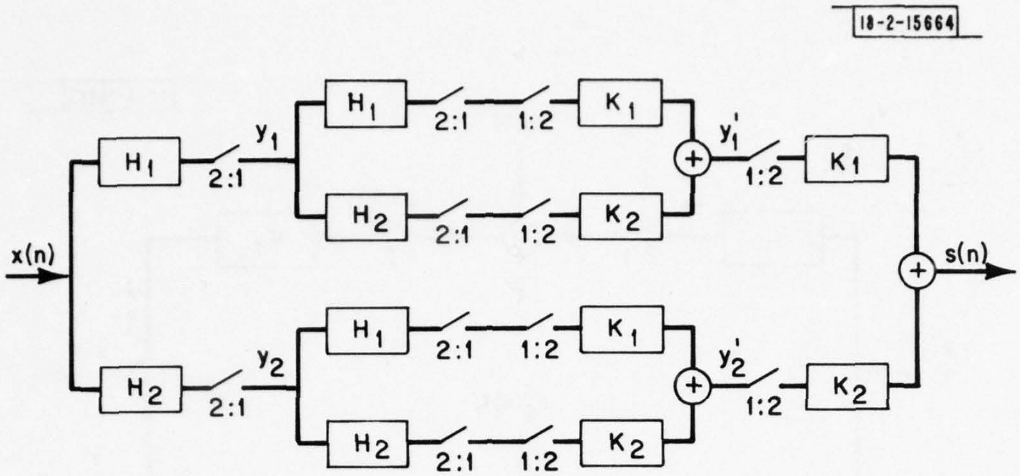


Fig. IV-5. Four-channel quadrature mirror-filter decomposition.

not affect the overall system output. It is clear that in that case, the output in Fig. IV-5 will be identical to the input.

If the only requirement placed on the filters in the 4-channel system is that they follow the relationships of Eq. (IV-3), the aliasing components will vanish. The system will behave as a linear filter with system response described as follows:

2-Channel System Function

$$G(z) = \frac{S(z)}{X(z)}$$

$$= \frac{1}{2} [H^2(z) - H^2(-z)] \quad . \tag{IV-7}$$

4-Channel System Function

$$G'(z) = \frac{1}{4} [H^2(z^2) - H^2(-z^2)] [H^2(z) - H^2(-z)]$$

$$= G(z^2) G(z) \quad . \tag{IV-8}$$

In the case that G(z) is an identity, then G'(z) will also be an identity system.

Decomposition into any number of channels that is a power of two can be performed by further extension of the basic scheme. In the same manner, it can also be shown that the aliasing components will vanish if the same relationships of the filters are maintained. The linear filter-system response will be a function of the basic filter that is used. For example, the 8-frequency band case will have the system function, G''(z):

8-Channel System Function

$$G''(z) = G'(z^2) G(z)$$

$$= G(z^4) G(z^2) G(z) \quad . \tag{IV-9}$$

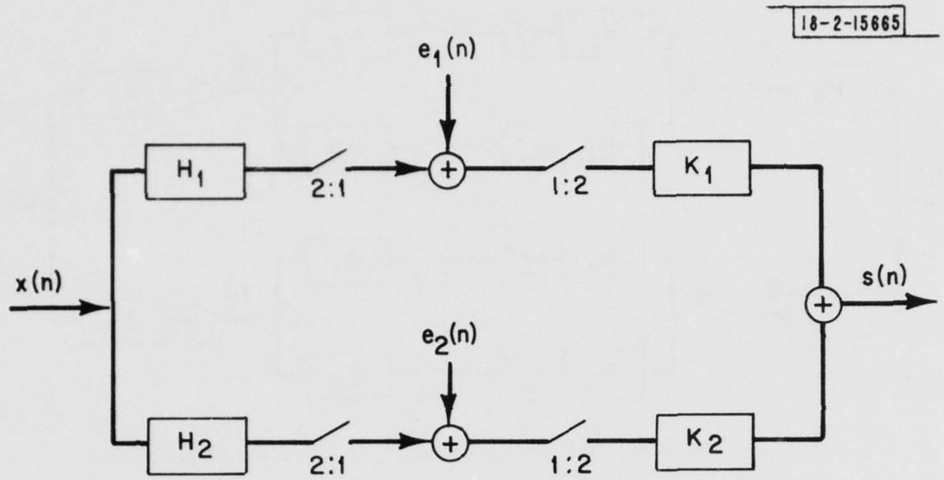


Fig. IV-6. Two-filter quadrature mirror-filter decomposition with quantization.

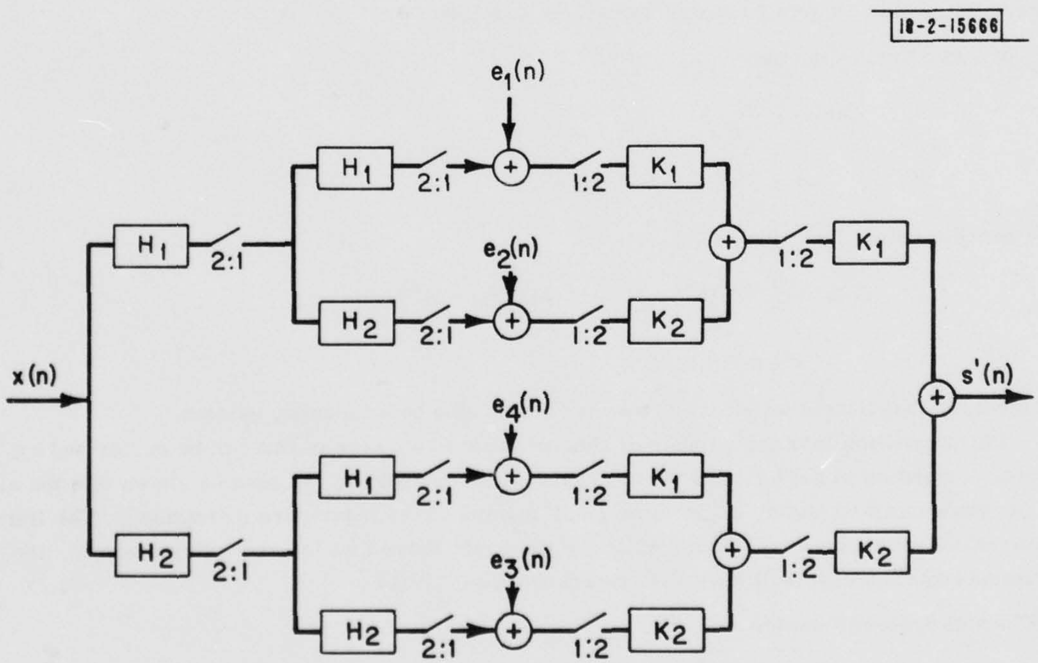


Fig. IV-7. Four-channel quadrature mirror-filter decomposition with quantization.

Although the decomposition into bands that are not of equal bandwidth does permit some aliasing at the output, this is minor and is discussed in Section IV-C-3 and in Appendix I.

2. Design of the Mirror Filters

It is shown in Appendix I that the class of filters that result in perfect reconstruction of the signal; i.e., an identity system, is not suitable for a practical system. This Section will explore the issues involved in the design of the prototype filter for use in the system.

There are several design constraints on the filters. With no quantization, coding, or processing other than the system filtering and resampling, the frequency response of the system should not introduce audible coloration to the signal. The reconstructed signal at the output must sound identical to the input signal. When quantization and coding are performed, the encoding error of a band should be contained to a spectral region close to that band.

The use of linear phase-FIR filters will result in a system function with no phase distortion, only linear phase components. The magnitude response of the basic half-band lowpass filter should be 3 dB down at $\pi/2$, the crossover with its mirror-image highpass filter. When the signal passes through a filter twice in each channel, the response of each channel will be 6 dB down at crossover. The two channels will than sum correctly at that frequency. By using filters with low-ripple passband response, the system-response ripple will be minimized.

When error is introduced to a channel through quantization of the signal, that noise is filtered and will result in noise at the output. Figure IV-6 shows the basic 2-channel, quadrature mirror-filter system with quantization, a nonlinear function modeled as an additive error signal. Since resampling, filtering, and summation are linear functions, this additive noise term will be a processed additive noise term at the output. Resampling affects the error signal by scaling the frequency axis. If the additive error is wideband, as can normally be assumed, this will not change its spectral extent. The error from quantization of a band in the two-band system will result in a wideband noise that is filtered once in that channel.

The multiband system is more complex as illustrated by the four-band system in Fig. IV-7. The error signal of each channel is resampled and filtered twice. The resampling operation after filtering scales the frequency axis in a manner that halves the filter passband bandwidth, sharpens the band edges, and inserts a replica of the filter shifted a distance of π from the original. Unfortunately, the band edges of every channel error signal are not defined by the sharpened filters. Figure IV-8 illustrates the filtering of the wideband noise generated by the quantization of each channel. The error-signal transition edges of channels 1 and 4 have been sharpened by a factor of two. Channels 2 and 3, however, have only one sharp edge. Hence, there will be some bands where the transition slope will only be as steep as the prototype filter, h(n).

As discussed in Chapter III, there is little masking at a distance of more than a critical band from the masker. Although the SNR in a band may only need to be on the order of 25 dB, the noise due to encoding of that band must be attenuated 80 dB or more at frequencies away from the band where there is no signal energy. Natural signals rarely have sharp discontinuities in their spectrum so that filter attenuation in the stop band next to the transition band need not be much more than 40 dB. However, the filter must have a stop-band attenuation that continues to increase away from the passband.

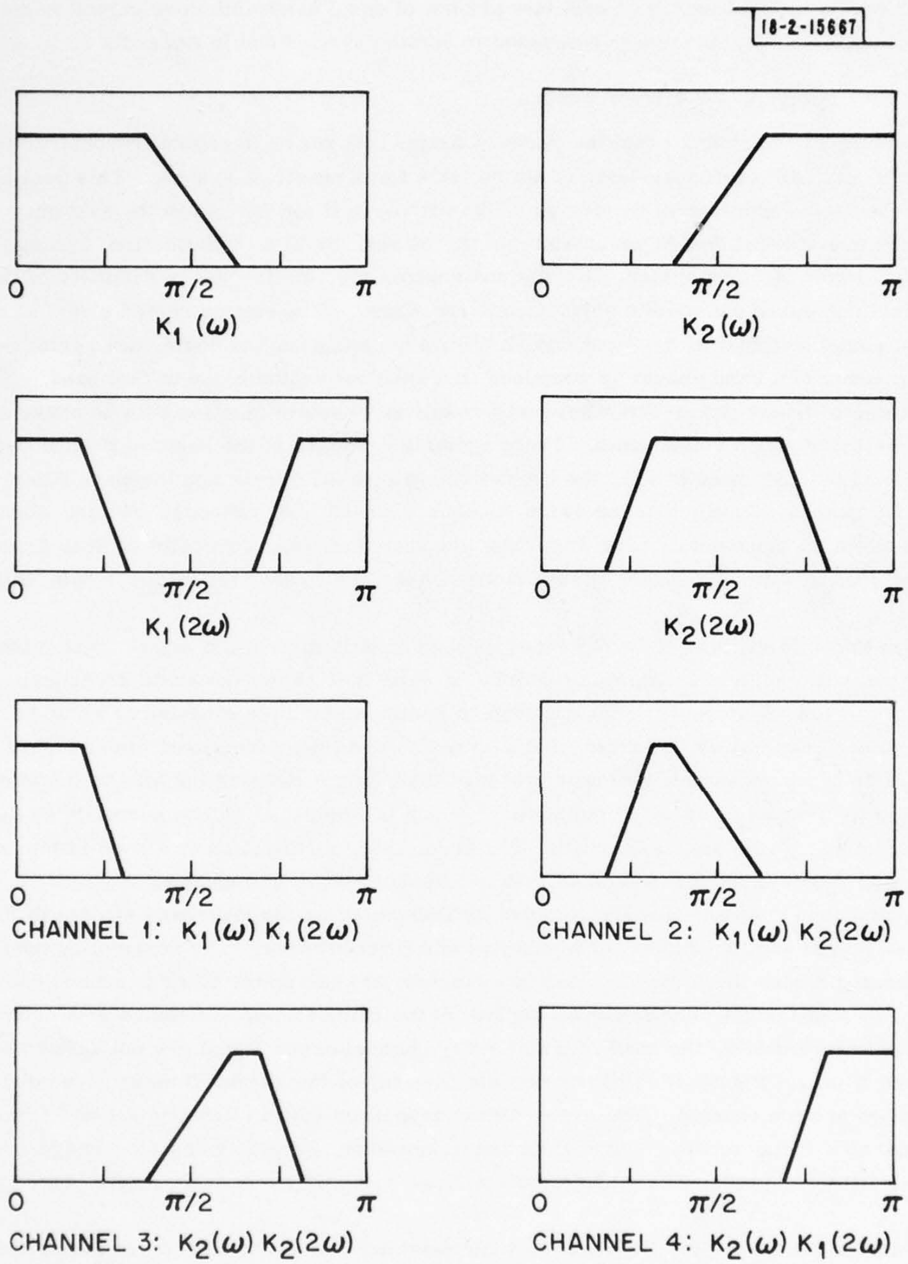


Fig. IV-8. Spectra of error at output produced by quantization of each channel in a four-channel quadrature mirror-filter system.

The filter chosen for use in the system is a Hanning windowed ideal bandpass filter of length 64. The width of the transition band is approximately $\pi/8$. A filter of higher order would have resulted in sharper transition bands at an increased computation cost. Evaluation of the system shows that this filter is adequate. The minimum stop-band attenuation is 44 dB. A Hanning window was chosen over other possible windows because attenuation in the stop-band increases at 18 dB per octave of frequency distance from the passband. This rapid increase ensures that little quantization noise is present far from the frequency of its band.

A method to take advantage of the sharpening of the filters of inner-band splitting would be to use filters of different lengths for different stages of decompositiion. This could result in great computational savings. In the eight-band system, for example, the innermost filters could be of length 32, the middle decomposition filters of length 64, and the final filters of length 128. The computation necessary would be equivalent to a system using filters of length 64, exclusively. All of the bands would then have the same slope, a factor of two sharper than several of the band-edge slopes in a system with all filters of length 64.

3. Unequal Bandwidth Filter Bank Design

The basic two-band, quadrature mirror-filter system and the extensions to four, eight, and larger number of channels are designed so that the bandwidths of each channel are identical. To match the bandwidths of each channel to the critical bandwidths desired, it is necessary to modify the scheme to permit decomposition into bands that are not of equal bandwidths.

The simplest solution is to divide the frequency range into many small bandwidth channels. The encoding of those channels that were to be of larger bandwidth would occur after partial reconstruction of the channels; i.e., the sum of several of the small bands. The analysis of equal bandwidth decomposition shows that no aliasing occurs in this scheme. It is very inefficient computationally in that the processing to divide bands that will not be used in their fully divided state is performed.

A more efficient method would be to use a partial tree structure where some branches are decomposed further than other baches and, therefore, result in smaller bandwidth channels. This is incorporated in Fig. IV-9, a three-channel system. It is clear that if it were possible

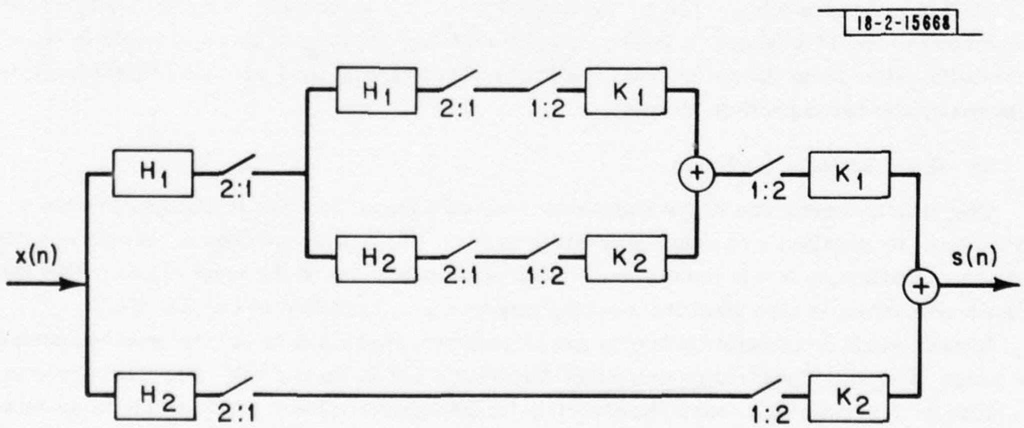


Fig. IV-9. Three-channel quadrature mirror-filter decomposition.

to design the basic two-channel decompostion as an identity system, its presence or absence in a branch would not affect the overall system function and any partial tree structure would be an identity system. It is shown in Appendix I, however, that it is not practical to use the filters necessary for the identity system. For the equal bandwidth decompositions in Section IV-C-1, it was shown that if the special relationships of the filters were maintained, the aliasing products would vanish. This result relies on the symmetry of the system such that the system equations factor properly. When a partial tree structure is used, some aliasing will remain. The output of the three-channel system is shown in Eq. (IV-10):

$$S(z) = \frac{1}{2}\,[H^2(z)\,G(z^2) - H^2(-z)]\,X(z) + \frac{1}{2}\,[G(z^2) - 1]\,[H(z)\,H(-z)]\,X(-z) \quad . \qquad \text{(IV-10)}$$

In terms of the Fourier transform:

$$S(\omega) = \frac{1}{2}\,[H^2(\omega)\,G(2\omega) - H^2(\omega + \pi)]\,X(\omega) + \frac{1}{2}\,[G(2\omega) - 1]\,[H(\omega)\,H(\omega + \pi)]\;X(\omega + \pi) \quad . \qquad \text{(IV-11)}$$

Of interest is the second term in Eqs. (IV-10) and (IV-11) representing the aliasing in the output. The aliasing will only appear in the frequency overlap of the filter h(n) and its mirror filter; i.e., frequencies where the product $H(\omega)\,H(\omega + \pi)$ is nonzero. This overlap is only in the region around $\pi/2$, the half-band splitting frequency. The aliasing is also scaled by $G(2\omega) - 1$. $G(\omega)$ is the linear system function of the basic two-band scheme and varies from unity only near $\pi/2$. The frequency-scaled $G(2\omega)$ varies from unity only near $\pi/4$ and $3\pi/4$. This scaling factor is very small at $\pi/2$. Hence, *the aliasing will still be minimal using the partial tree structure for* unequal bandwidth decomposition. To summarize the analysis in terms of design constraints, there will be little aliasing if the two-band, linear-system-function, $G(\omega)$, has ripple only at a frequency of $\pi/2$ and $H(\omega)$ is very nearly zero at a frequency of $3\pi/4$. The Hanning-window-design filters meet this requirement.

Using the partial tree, unequal-bandwidth, channel-decomposition technique, the structure of Fig. IV-1 is implemented. The actual bandwidths of the 24-channel, 15-kHz, audio-encoding system and of the 17-channel, 4.1-kHz, speech-encoding system are given in Table IV-1 . These bandwidths differ from the critical bandwidths as shown in Fig. III-4 because of the constraints imposed by the decomposition technique.

D. Quantization Algorithms

The principal requisite of the quantization scheme is that it adapt to changes in channel signal level to maintain a constant percentage error. For a given number of bits for quantization, the quantization levels must allow for the amplitude peaks of the input signal. The SNR in each channel must also meet the masking requirements specified in Chapter III.

Instantaneous companding schemes use logarithmic spacing of quantizer levels to permit the large, dynamic, input range necessary for speech and audio signals. The RMS error in encoding each sample is a constant percentage of the sample value. Each sample is quantized independently.

TABLE IV-1
CHANNEL BANDWIDTHS OF DIGITAL ENCODER

Channel Number	4.1-kHz Speech Encoding System Frequency Range (Hz)	4.1-kHz Speech Encoding System Bandwidth (Hz)	15-kHz Audio Encoding System Frequency Range (Hz)	15-kHz Audio Encoding System Bandwidth (Hz)
0	0 – 129	129	0 – 117	117
1	129 – 258	129	117 – 234	117
2	258 – 387	129	234 – 351	117
3	387 – 516	129	351 – 468	117
4	516 – 645	129	468 – 585	117
5	645 – 774	129	585 – 703	117
6	774 – 903	129	703 – 820	117
7	903 – 1033	129	820 – 937	117
8	1033 – 1291	258	937 – 1171	234
9	1291 – 1549	258	1171 – 1406	234
10	1549 – 1807	258	1406 – 1640	234
11	1807 – 2066	258	1640 – 1875	234
12	2066 – 2324	258	1875 – 2109	234
13	2324 – 2582	258	2109 – 2343	234
14	2582 – 3099	516	2343 – 2812	468
15	3099 – 3615	516	2812 – 3281	468
16	3615 – 4132	516	3281 – 3750	468
17			3750 – 4687	937
18			4687 – 5625	937
19			5625 – 6562	937
20			6562 – 7500	937
21			7500 – 9375	1875
22			9375 – 11250	1875
23			11250 – 15000	3750

Speech and natural audio signals do not vary instantaneously in level. The average power is usually constrained by attack transients with time constants greater than 10 ms, and decay transients with time constants greater than 100 ms. Adaptive quantization systems take advantage of this slow variation of the short-time average power by specifying a normalization gain factor at a much lower rate than the sampling of the signal. This yields a lower bit rate than the instantaneous companders. Many adaptive strategies, however, tend to overload when presented with a sharp attack transient. For a short time, this overload can produce a very large error that would be audible. Block-companding adaptive PCM was chosen because of its immunity to overload. Sometimes known as block-floating-point encoding, the scheme uses a quantized block maximum magnitude to normalize all samples in the block. Each sample is then quantized by linear PCM. (An optimized distribution of quantization levels may be used.[39] The decrease in RMS error will depend on how well the sample value probability distribution is estimated and the number of quantization levels in use.) By quantizing the block maximum to a level larger than the block maximum, overload is avoided.

Since the samples in a block are all quantized using the same normalization, the RMS error is constant over the block. If the block length is too long, an increase in signal level near the end of a block will cause a proportional increase in error energy throughout the block that will not be masked by the backward masking (premasking) effect of the signal. As the temporal extent of backward masking is much less than forward masking (postmasking), this is the limiting case. The block lengths were chosen to be inversely proportional to the channel bandwidth as is the time window in the peripheral auditory system used for the short-time spectral analysis on the basilar membrane.[13] Due to the lack of nonsimultaneous masking data as a function of stimuli frequency, it is not clear whether the temporal integration for masking follows the same proportionality. Block lengths of eight samples in each channel were chosen for the encoder. These lengths in seconds for each channel are given in Table IV-2. In any case, it is likely that the block lengths can be made longer in the high-frequency channels since virtually all naturally produced audio signals have attack times much longer than the block lengths of these channels.

TABLE IV-2
BLOCK LENGTHS OF EACH CHANNEL OF DIGITAL ENCODER

Channel Number	4.1-kHz Speech Encoding System		15-kHz Audio Encoding System	
	Frequency Range (Hz)	Block Length (ms)	Frequency Range (Hz)	Block Length (ms)
0	0 – 129	31.0	0 – 117	34.2
1	129 – 258	31.0	117 – 234	34.2
2	258 – 387	31.0	234 – 351	34.2
3	387 – 516	31.0	351 – 468	34.2
4	516 – 645	31.0	468 – 585	34.2
5	645 – 774	31.0	585 – 703	34.2
6	774 – 903	31.0	703 – 820	34.2
7	903 – 1033	31.0	820 – 937	34.2
8	1033 – 1291	15.5	937 – 1171	17.1
9	1291 – 1549	15.5	1171 – 1406	17.1
10	1549 – 1807	15.5	1406 – 1640	17.1
11	1807 – 2066	15.5	1640 – 1875	17.1
12	2066 – 2324	15.5	1875 – 2109	17.1
13	2324 – 2582	15.5	2109 – 2343	17.1
14	2582 – 3099	7.8	2343 – 2812	8.52
15	3099 – 3615	7.8	2812 – 3281	8.52
16	3615 – 4132	7.8	3281 – 3750	8.52
17			3750 – 4687	4.26
18			4687 – 5625	4.26
19			5625 – 6562	4.26
20			6562 – 7500	4.26
21			7500 – 9375	2.13
22			9375 – 11250	2.13
23			11250 – 15000	1.07

V. EVALUATION

A. Introduction

The digital encoding system is designed so that the reconstructed output signal should sound identical to the input signal if the SNR requirements are met in each channel. The experiments to estimate those requirements were performed, however, using tone and narrowband noise stimuli. It can not be expected that the perception of the complex sounds that are present in speech and audio signals is exactly the same as for the primitive stimuli. The perception of speech sounds, for example, is related to the understanding of speech and the generation of speech.[40] It was not known whether the masking under these conditions is greater than, less than, or the same as the masking present in the simple stimuli experiments.

The system was evaluated using speech signals with a 4.1-kHz bandwidth and music selections of 15-kHz bandwidth. Experiments were performed to determine the level of quantization error and the bit rate at the differential threshold of the encoding error.

For many applications, perfect speech reproduction is not necessary. For these communication systems, it is desired to have highly intelligible, pleasant and natural sounding speech at lower bit rates for use on less expensive, lower-capacity channels. To meet these demands, a 17-channel, 3.2-kHz version of the system was evaluated at a bit rate of 16 kbps.

B. High-Quality Speech Encoding

To test the performance of the system for speech encoding, a 17-band, 4.1-kHz configuration was used. The goal was to determine the lowest bit rate and the corresponding quantization bit distribution to encode speech at which the encoding error is just detectable. The error intensity at which the encoding error is just detectable is the masked or differential threshold of the encoding error in the presence of the signal at that signal level. Controlled experiments were performed to determine this threshold.

The source material was 24 phonetically balanced sentences, each approximately two seconds in length. Four speakers, two male and two female, were recorded with an Electrovoice 667 microphone in a soundproof room in digital format through a 16-bit A/D converter. For testing, the speech was played back directly from digital storage through a 16-bit D/A converter and Stax SR-X electrostatic headphones in the soundproof room.

Experiments consisted of two-interval, two-alternative, forced-choice (2I2AFC) trials to compare the original, unprocessed sources to the encoded, processed versions of the same sentence. Each interval of a trial was chosen randomly and independently of the other interval to have either the original or processed version of a sentence. Thus, each of the four permutations of original and processed versions of the same sentence, same speaker, were likely to occur.

Subjects were asked to judge the two intervals as being identical or not identical. Any audible difference, noise, distortion, coloration, etc., was valid for a "not identical" judgment. Subjects were told a priori that the probabilities of the intervals being the same or being different were equal.

If a subject could not discriminate any audible differences between the original and processed sentences in any trial, the probability of a correct response would be 50-percent, random guessing. Just-detectable degradation would result in a probability of 75-percent correct, half way between 50-percent chance and 100-percent perfect detection. Thus, the error present at this score is defined as the differential sensory threshold of the encoding degradation (JND).[41]

The test sessions consisted of 24 training trials with feedback followed by 48 trials that were scored. Seven sets of system parameters, each representing a quantization bit distribution and a resultant bit rate, were used. Four to six subjects were tested with each set of system parameters. The results of the experiments are presented in Table V-1 and Fig. V-1. A statistical analysis of the experimental procedure and explanation of the results can be found in Appendix II.

The performance of the system as a function of the quantization bit distribution shows that the SNR in the high-frequency channels can be less than that needed by the mid-frequency channels. System E from Table V-1 with a bit rate of 34.4 kbps yields an average subject performance score of 72-percent correct, just under the JND threshold. Thus, this system can reproduce speech without audible degradation. It is difficult to compare this performance with other digital encoders because few systems have been evaluated at this high performance level. It can be speculated that the introduction of a highly optimized scheme, similar to that of adaptive transform coding (ATC), could further reduce the bit rate and yield comparable results.

TABLE V-1

QUANTIZATION BIT DISTRIBUTION AND EXPERIMENTAL RESULTS OF SPEECH ENCODING SYSTEM

Channel Number	Frequency Range (Hz)	Quantization Bit Distribution						
		A	B	C	D	E	F	G
0	0 – 129	4	3	4	3	3	3	3
1	129 – 258	4	4	4	4	4	3	3
2	258 – 387	4	4	4	4	4	4	4
3	387 – 516	4	4	4	4	4	4	4
4	516 – 645	4	4	4	4	4	4	4
5	645 – 774	4	4	4	4	4	4	4
6	774 – 903	4	4	4	4	4	4	4
7	903 – 1033	4	4	4	4	4	4	4
8	1033 – 1291	4	4	4	4	4	4	4
9	1291 – 1549	4	4	4	4	4	4	4
10	1549 – 1807	4	4	4	4	4	4	4
11	1807 – 2066	4	4	4	4	4	4	4
12	2066 – 2324	4	4	4	4	4	4	4
13	2324 – 2582	4	4	4	4	3	4	3
14	2582 – 3099	4	4	4	3	3	3	3
15	3099 – 3615	4	4	3	3	3	3	3
16	3615 – 4132	4	3	3	3	3	3	3
Date Rate (kbps):		38.3	36.9	36.2	34.9	34.4	34.6	34.1
Percent Correct:		63	67	68	64	72	85	94

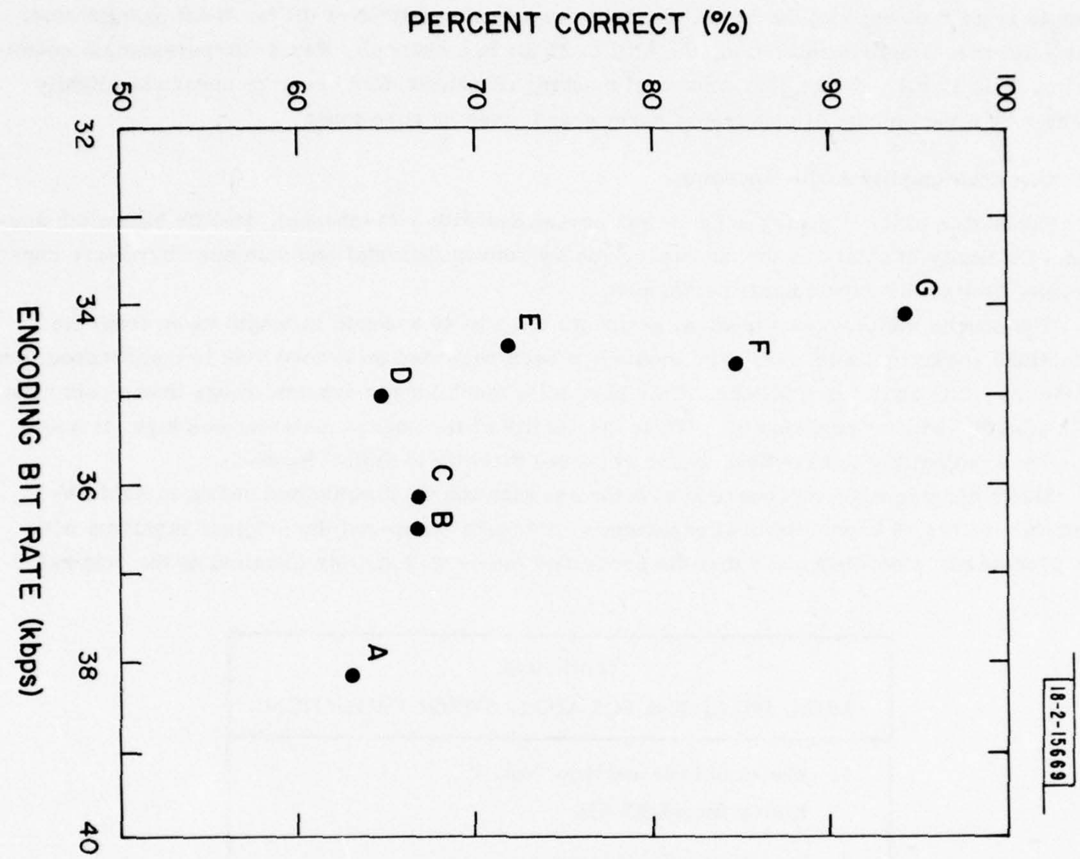


Fig. V-1. Relation of the bit rate to the experimental results.

These results can be compared to the results of the masking experiment presented in Chapter III. Those results, shown in Table III-1 and Fig. III-6, indicated that the masked threshold of a narrowband of noise masked by a sinusoid at 70-dB SPL centered in the noise is between 42- and 52-dB SPL. This is an SNR at the masked threshold of 18 to 28 dB dependent on the frequency of the stimuli. The SNR present in each channel of the encoding system for System E can be estimated from the quantization algorithm. The encoding degradation for the quantization bit distribution of System E has been found to be below the masked threshold.

Each channel is quantized with a block-companding APCM scheme. The adaption is performed by using a scale factor to normalize a block of channel samples. Each block uses one of 24 scale factors, each 6-dB apart. The scale factor is chosen so that it was always greater than the largest magnitude of any sample in the block. Thus, it averages 3-dB greater than the largest sample. After normalization, the samples in the block are quantized by linear PCM. Assuming that the companding can adapt quickly enough such that the quantization range up to the largest sample is used throughout the block, the SNR is approximately 6N-1 dB for N-bit quantization. For 4-bit-per-sample quantization, the SNR is 23 dB in a channel. For 3-bit-per-sample quantization, it is 17 dB. Hence, the amount of masking of quantization error by speech is slightly greater than the amount of masking of narrowband noise by pure tones.

C. High-Quality Audio Encoding

Evaluation of the encoding scheme was performed with a 24-channel, 15-kHz bandwidth system. Difficulty in obtaining the necessary quality source material and computer hardware constraints limited the experiments performed.

The source material was musical segments of 25 to 40 seconds in length taken from the selections shown in Table V-2. The music had been recorded on record disk in compressed form by the use of an analog compander. This permits a much larger dynamic range than would have been possible without companding. While the quality of the source material was high, it would have been preferable to have used music recorded directly in digital format.

Music segments were processed with the quantization bit distribution shown in Table V-3, a bit rate of 123.75 kbps. Several experienced listeners compared the original segments with the processed. Consensus was that the processed music was audibly identical to the original.

TABLE V-2

MUSIC SELECTIONS FOR AUDIO SYSTEM EVALUATION

1. Masters of Flute and Harp, Vol. 1, Klavier Records KS-556
2. Stan Kenton Plays Chicago, Creative World Records ST-1072
3. Rags and Other American Things, Eastern Brass Quintet, Klavier Records KS-539
4. The Heralds of Love, Klavier Records KS-559
5. Bach: Praeludim from Partita #1 in B flat, Klavier Records KS-524
6. St. Saens Organ Symphony #3 in C minor, Opus 78, The City of Birmingham Orchestra, Klavier Records KS-526

TABLE V-3

QUANTIZATION BIT DISTRIBUTION AND EXPERIMENTAL RESULTS OF AUDIO ENCODING SYSTEM

Channel Number	Frequency Range (Hz)	Quantization Bit Distribution
0	0 – 117	4
1	117 – 234	4
2	234 – 351	4
3	351 – 468	4
4	468 – 585	4
5	585 – 703	4
6	703 – 820	4
7	820 – 937	4
8	937 – 1,171	4
9	1,171 – 1,406	4
10	1,406 – 1,640	4
11	1,640 – 1,875	4
12	1,875 – 2,109	4
13	2,109 – 2,343	4
14	2,343 – 2,812	4
15	2,812 – 3,281	4
16	3,281 – 3,750	4
17	3,750 – 4,687	4
18	4,687 – 5,625	4
19	5,625 – 6,562	4
20	6,562 – 7,500	4
21	7,500 – 9,375	3
22	9,375 – 11,250	3
23	11,250 – 15,000	3

18-2-15670

(a)

(b)

(c)

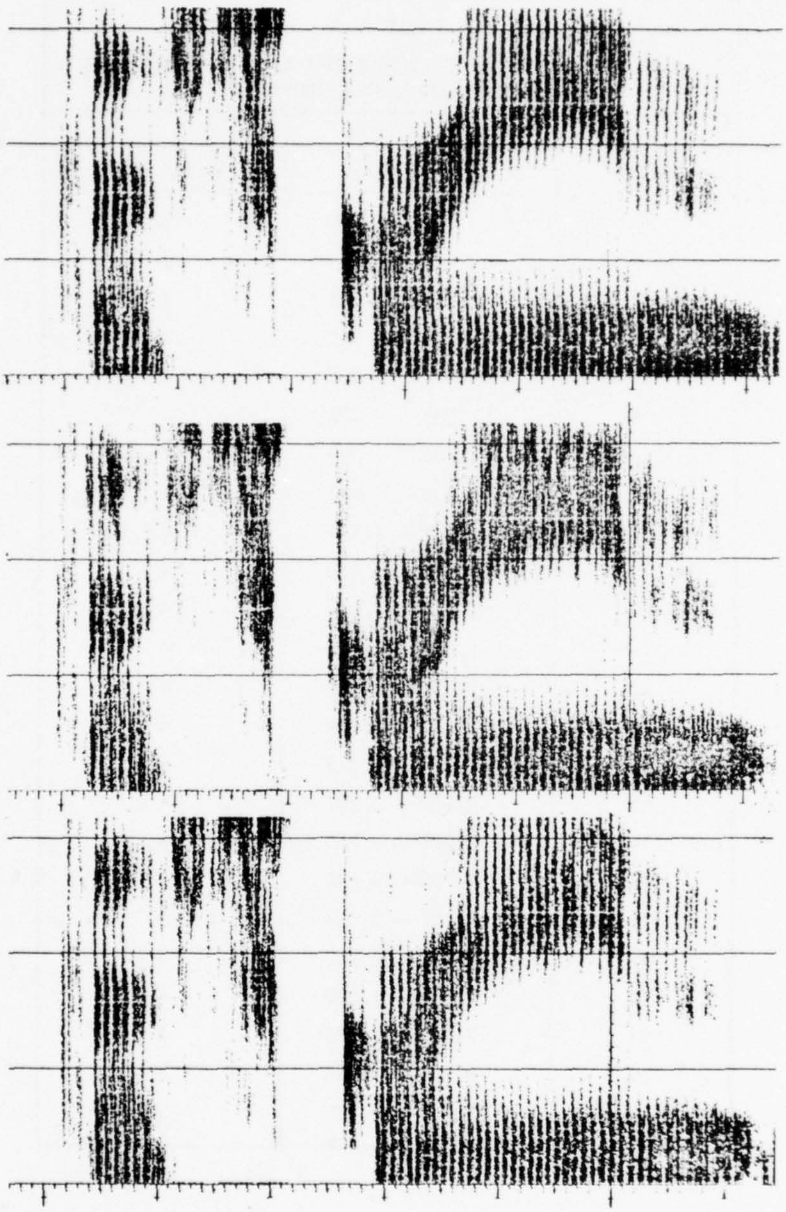


Fig. V-2(a-c). Spectrograms showing speech deformation, (a) original speech, (b) block-companded APCM speech, (c) smoothed magnitude APCM speech.

D. Toll-Quality Speech Encoding

For many communication systems, it is not necessary to have perfect reproduction of the speech signals. A lower quality is acceptable as long as it is highly intelligible, pleasant and natural sounding. Quantifying the acceptable level is difficult since it is actual-system-user acceptance that determines the necessary fidelity. In the literature, this quality is referred to as toll or telephone quality.

The encoding system developed here was designed using the results of psychoacoustic experiments to achieve an inaudible encoding error. It is conjectured that when the level of encoding error is too large for inaudibility, the system will degrade the speech in a pleasant manner. The spectral and temporal shaping of this degradation would be such that a larger noise level and, therefore, a lower bit rate, would be acceptable to a listener than for wideband waveform coding schemes such as ADPCM. The achievement of toll quality speech transmission at 16 kbps with an ATC system supports this speculation.[17]

For evaluation at 16 kbps, a 17-channel, 3.2-kHz system was implemented. Speech processed by this encoder, however, has a rough quality. To correct this particularly annoying form of degradation, it is necessary to analyze the assumptions made in the design of the encoder. The block-companding, adaptive-PCM-coding strategy was based on nonsimultaneous masking results. At a bit rate of 16 kbps, two bits is the average quantization per sample, not including the quantization of the block magnitude. The SNR as was developed in Section V-B is approximately 11 dB. Clearly, the quantization error will not be masked. The block-coding method causes the error to be of constant level for each entire block. In speech segments where there are large amplitude transitions, this noise will tend to extend the speech abruptly to the block edges. The noise energy can be greater than the speech energy right before an attack transient back to the beginning of the companding block. This deformation of the speech can be seen on the speech spectrograms in Fig. V-2a and b.

To eliminate this source of distortion, the companding scheme was modified as shown in Fig. V-3. The main purpose of this modification is so that the adapting magnitude is now a smooth function. The magnitude of the samples is filtered and then sampled at a 50-Hz rate. The adaptation time of the quantizer is 20 ms. Since the quantization error is also scaled by the magnitude function, it is now a smoothly varying noise signal in each channel. The system is no longer immune to overload on sharp transients with rise times less than 20 ms. The problem of overload is reduced by scaling the quantized magnitude function to permit peak samples several dB larger. This, of course, increases the quantization error level by the same amount. A compromise value was found to optimize the perceptual quality of the output. The magnitude is quantized to the nearest larger level of a set of levels spaced in 3-dB increments. It is then scaled up by an additional 3 dB. The result can be seen in the spectrogram (Fig. V-2c).

At a bit rate of 15.8 kbps, several trained listeners compared the system to a 16-kbps adaptive-predictive-coding (APC) system. High-quality speech as well as sentences with varied levels of background noise were used as input signals. The quality of the output speech of both systems were similar and both were of toll quality. The robustness, the ability to reproduce speech that is imbedded in background noise, of the system also compared well with the APC system.

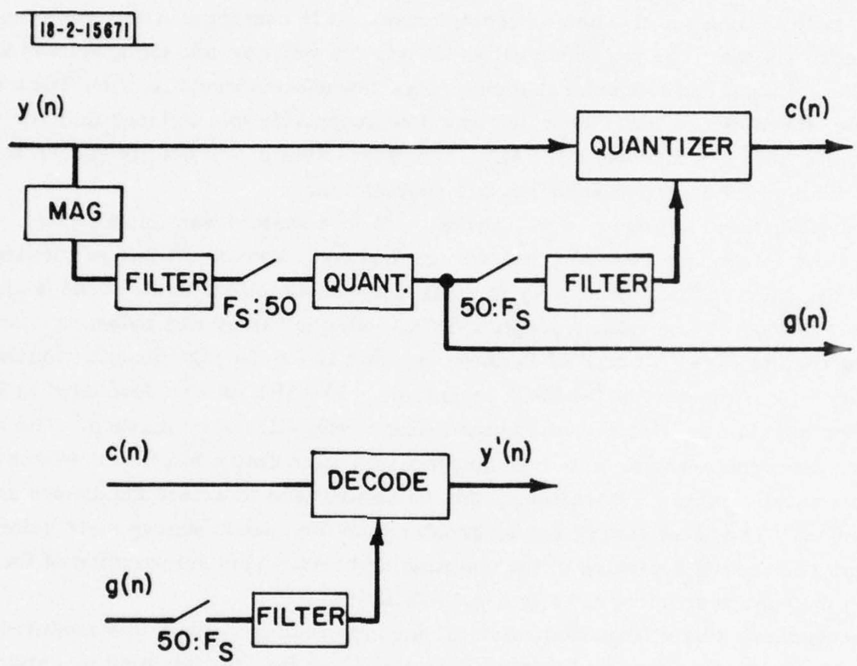


Fig. V-3. Quantization scheme for 16-kbps speech encoding.

VI. SUMMARY AND TOPICS FOR FURTHER RESEARCH

A. Summary

In this report, a digital encoder for speech and audio signals has been described. The technique of transformation of the signal to a domain where quantization is better matched to audition was developed based on the results of psychoacoustic experiments. By exploiting the limited detection ability of the auditory system as determined by masking experiments, the system achieves performance that is comparable to or better than other encoders at the same encoding bit rate. The required spectral and temporal shaping of the error is accomplished by use of a multichannel system with adaptive quantization in each channel. The channel bandwidths and quantization adaptation properties are related to the masking results.

Efficient decomposition of the input signal into critical bandwidth channels was performed by a quadrature mirror-filter scheme. The method was developed from a basic two-channel decomposition described in the literature.[42] An extensive discussion of the issues of design and implementation of the quadrature mirror filters including practical design of the filters and modification for a system with unequal bandwidth channels was presented.

Evaluation shows that the system achieves encoding that is audibly identical to the original signal at bit rates lower than were accomplished previously.

B. Topics for Further Research

The design of digital encoders for speech and audio signals that account for the processing and limitations of the auditory system is a new field. The system developed here demonstrates the feasibility of such systems and provides a framework for further research. Although the signal processing necessary for the implementation of this system requires computation that forbids a real-time implementation with the facilities available, the technology is being developed to allow it. The quadrature mirror-filtering technique is well suited to charge-coupled device (CCD), sampled-data filtering. Within a few years, it should be possible to implement multichannel encoding systems inexpensively as real-time systems.

There are several areas for further development of this digital encoder. The psychoacoustic results used to determine the perceptual requirements for the encoding were taken primarily from the literature. The experiments used simple stimuli such as tones, clicks, and white noise. It is difficult to relate these results to the perception of complex signals such as speech and audio. A better understanding of the perceptual requirements for speech and audio signal processing systems is necessary for further system development.

The SNR required in each channel of the encoding system for the encoding degradation to be at the threshold of detectability is slightly less than the SNR necessary for critical bandwidth noise to be masked by a sinusoid centered in the noise. Although the critical band concept appears in many psychoacoustic phenomena and in the physiology of the auditory system, it is not certain that it applies directly for this system. If the masking by a sinusoid is concentrated in a bandwidth smaller than a critical bandwidth, then a system with smaller bandwidth channels may perform better than the critical bandwidth channel system. On the other hand, if the spread of the masking is to a bandwidth greater than a critical bandwidth, a system with fewer channels, each of a larger bandwidth, may perform as well as the critical bandwidth channel system. Fewer channels would require less computation and, therefore, less costly implementation. The results of an experiment comparing the amount of masking of various bandwidths of narrowband noise by sinusoids would permit further optimization of the channel bandwidths of the encoding system.

The results of the encoding system at an encoding rate of 16 kbps, indicate that encoding degradation may be more audible during certain speech sounds than others. In particular, listeners appear to be very sensitive to segments containing onset transients. To confirm this, an examination of the amount of masking by different speech sounds of various noise signals is necessary. The results could be used to design an adaptive, quantization bit distribution system that could allocate quantization bits temporally and spectrally and, therefore, control the SNR in each channel at all times, according to the particular requirements of each speech sound.

Another area is development of the signal processing techniques. Introduction in multichannel encoders of algorithms to implement processing such as variable-length coding and adaptive bit distribution, signal prediction, and use of masking between channels, could result in further bit rate reduction. The development of these techniques depends on the understanding of the perceptual requirements for encoding systems.

REFERENCES

1. Rabiner, L. R. and Johnson, J. A., "Perceptual Evaluation of the Effects of Dither on Low Bit Rate PCM Systems," Bell System Tech. J. 51, 1487-1494 (1972).
2. Cummiskey, P.; Jayant, N. S.; and Flanagan, J. L., "Adaptive Quantization Differential PCM Coding of Speech," Bell System Tech. J. 52, 1105-1118 (1973).
3. Wozencraft, J. M. and Jacobs, I. M., Principles of Communications Engineering (Wiley, New York, 1965).
4. Bennett, W. R., "Spectra of Quantized Signals," Bell System Tech. J. 27, 446-472 (1948).
5. Rabiner, L. R. and Schafer, R. W., Digital Processing of Speech Signals (Prentice Hall, Englewood Cliffs, N.J., 1978).
6. Smith, B., "Instantaneous Companding of Quantized Signals," Bell System Tech. J. 36, 653-709 (1957).
7. Hessenmuller, H., "The Transmission of Broadcast Programs in a Digital Integrated Network," IEEE AU-21, No. 1, 17-20 (1973).
8. Osborne, D. W., "Digital Sound Signals: Further Investigation of Instantaneous and Other Rapidly Companding Systems," British Broadcasting Corp. RD-1972/31 (1972).
9. Osborne, D. W. and Croll, M. G., "Digital Sound Signals: Bit-Rate Reduction Using an Experimental Digital Compander," British Broadcasting Corp. RD-1973/41 (1973).
10. Atal, B. S. and Schroeder, M. R., "Adaptive Predictive Coding of Speech Signals," Bell System Tech. J. 49, 1973-1986 (1970).
11. Noll, P., "A Comparative Study of Various Schemes for Speech Encoding," Bell System Tech. J. 54, 611-632 (1974).
12. Makhoul, J. I. and Wolf, J. J., "Linear Prediction and the Spectral Analysis of Speech," Bolt Beranek & Newman Report 2304, 89-138 (1972).
13. Flanagan, J. L., Speech Analysis Synthesis and Perception (Springer-Verlag, New York, 1972).
14. Young, I. M. and Wenner, C. H., "Masking of White Noise by Pure Tone, Frequency Modulated Tone, and Narrow-Band Noise," J. Acoust. Soc. Am. 41, 700-706 (1967).
15. Crochiere, R. E.; Webber, S. A.; and Flanagan, J. L., "Digital Coding of Speech in Sub-bands," Bell System Tech. J. 55, 1069-1085 (1976).
16. Zelinski, R. and Noll, P., "Adaptive Transform Coding of Speech Signals," IEEE Trans. Acoust., Speech, and Sig. Proc. 25, 299-309 (1977).
17. Flanagan, J. L.; Schroeder, M. R.; Atal, B. S.; Crochiere, R. E.; Jayant, N. S.; and Tribolet, J. M., "Speech Coding," IEEE COM-27, 710-736 (1979).
18. Chew, J. R., "Pulse-code Modulation for High Quality Sound Signal Distribution: Feasibility of a Multiplex System," British Broadcasting Corp. RD-1970/34 (1970).
19. Croll, M. G., "Pulse Code Modulation for High Quality Sound Distribution: Quantizing Distortion at Very Low Signal Levels," British Broadcasting Corp. RD-1970/18 (1970).
20. Croll, M. G.; Moffat, M. E. B.; and Osborne, D. W., "'Nearly Instantaneous' Compandor for Transmitting Six Sound-Programme Signals in a 2.048 Mbit/s Multiplex," Electronics Letters 9, 14, 298-300 (1973).
21. ISO Recommendation R226, "Normal Equal Loudness Contours for Pure Tones and Normal Threshold of Hearing Under Free Field Listening Conditions," International Standards Organization (1961).

22. Licklider, J. C. R., "Basic Correlates of the Auditory Stimulus," Handbook of Experimental Psychology (Wiley, New York, 1951).

23. Fletcher, H., Speech and Hearing (Van Nostrand, New York, 1929).

24. Egan, J. P. and Hake, H. W., "On the Masking Pattern of Simple Auditory Stimulus," J. Acoust. Soc. Am. 22, 622-630 (1950).

25. Jeffress, L. A., "Masking," Foundations of Modern Auditory Theory, 87-114 (Academic Press, New York, 1970).

26. Scharf, B., "Complex Sounds and Critical Bands," Psychol. Bul. 58, 205-217 (1961).

27. Scharf, B., "Critical Bands," Foundations of Modern Auditory Theory, 159-202 (Academic Press, New York, 1970).

28. Zwicker, E., "Subdivision of the Audible Frequency Range into Critical Bands," J. Acoust. Soc. Am. 33, 248 (1961).

29. Hawkins, J. E., Jr. and Stevens, S. S., "The Masking of Pure Tones and of Speech by White Noise," J. Acoust. Soc. Am. 22, 6-13 (1950).

30. Young, I. M., "Effects of Pure-Tone Masking on Low-Pass- and High-Pass-Filtered Noise," J. Acoust. Soc. Am. 45, 1206-1209 (1969).

31. Lee, F. F. and Lipschutz, D., "Floating Point Encoding for Transcription of High Fidelity Audio Signals," Audio Eng. Soc. 55th Convention, preprint 1196 (1976).

32. Fastl, H., "Temporal Masking Effects: II. Critical Band Noise Masker," Acoustica 36, 317-330 (1977).

33. Schafer, R. W. and Rabiner, L. R., "Design of Digital Filter Banks for Speech Analysis," Bell System Tech. J. 50, 3097-3115 (1971).

34. Schafer, R. W. and Rabiner, L. R., "Design and Simulation of a Speech Analysis-Synthesis System Based on Short-Time Fourier Analysis," IEEE AU-21, 3, 165-174 (1973).

35. Portnoff, M. R., "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform," IEEE Acoustics, Speech, and Signal Processing ASSP-24, 243-248 (1976).

36. Crochiere, R. E. and Sambur, R. M., "A Variable Band Coding Scheme for Speech Encoding at 4.8 kb/s," IEEE Acoustics, Speech, and Signal Processing Conference, Hartford, Connecticut (1977).

37. Esteban, D. and Galand, C., "Application of Quadrature Mirror Filters to Split Band Voice Coding Schemes," Proc. IEEE Acoustics, Speech, and Signal Processing Conference, 191-195, Hartford, Connecticut (1977).

38. Croisier, A.; Esteban, D.; and Galand, C., "Perfect Channel Splitting by Use of Interpolation/Decimation/Tree Decomposition Techniques," Proc. First Intl. Conf. on Info. Sci. and Systems, 443-446, Patras, Greece, (1976).

39. Max, J., "Quantizing for Minimum Distortion," IRE Trans. Info. Theory, IT-6, 7-12 (1960).

40. Stevens, K. N. and House, A. S., "Speech Perception," Foundations of Modern Auditory Theory, 2, 3-62 (Academic Press, New York, 1970).

41. Green, D. M. and Swets, J. A., Signal Detection Theory and Psychophysics (Wiley, New York, 1966).

42. Esteban, D. and Galand, C., "Parallel Approach to Quasi-Perfect Decomposition of Speech in Sub-Bands," Ninth Intl. Cong. on Acoustics, Madrid, Spain (1977).

43. Rabiner, L. R. and Gold, B., Theory and Applications of Digital Signal Processing (Prentice Hall, Englewood Cliffs, N.J., 1975).

ADDITIONAL REFERENCES

Berouti, M. and Makhoul, J., "High Quality Adaptive Predictive Coding of Speech," Proc. IEEE Acoustics, Speech, and Signal Proc. Conf., Tulsa, Oklahoma, 303-306 (1978)

Bilger, R. C., "A Revised Critical-Band Hypothesis," Univ. of Pittsburgh (to be published).

Bilger, R. C., "Additivity of Different Types of Masking," J. Acoust. Soc. Am. 31, 1107-1109 (1959).

Bilger, R. C., "Masking of Tones by Bands of Noise," J. Acoust. Soc. Am. 28, 623-620 (1956).

Blesser, B. A., "Audio Dynamic Range Compression for Minimum Perceived Distortion," IEEE AU-17, 22-32 (1969).

Blesser, B. A. and Kates, J. M., "Digital Processing in Audio Signals," Applications of Digital Signal Processing, 29-116 (Prentice-Hall, Englewood Cliffs, N.J., 1978).

Deatherage, B. H., Davis, H., and Eldredge, D., "Physiological Evidence for the Masking of Low Frequencies by High," J. Acoust. Soc. Am. 29, 132-137 (1957).

Diericks, K. J. and Jeffress, L. A., "Interaural Phase and the Absolute Threshold for Tone," J. Acoust. Soc. Am. 34, 981-984 (1962).

Duifhuis, H., "Consequences of Peripheral Frequency Selectivity for Nonsimultaneous Masking," J. Acoust. Soc. Am. 54, 1471-1488 (1973).

Duifhuis, H., "Audibility of High Harmonics in a Periodic Pulse," J. Acoust. Soc. Am. 48, 888-893 (1970).

Duifhuis, H., "Audibility of High Harmonics in a Periodic Pulse, II: Time Effect," J. Acoust. Soc. Am. 49, 1155-1162 (1971).

Elliott, L. L., "Backward Masking: Monotic and Dichotic Conditions," J. Acoust. Soc. Am. 34, 1108-1115 (1967).

Elliott, L. L., "Backward and Forward Masking of Probe Tones of Different Frequencies," J. Acoust. Soc. Am. 34, 1116-1117 (1967).

Esteban, D., "Optimum N-Bit PCM Block Quantizing," Proc. First Intl. Conf. on Info. Sci. and Systems, Patras, Greece, 728-733 (1976).

Esteban, D. and Galand, C., "32-KBPS CCITT Compatible Split-Band Coding Scheme," Proc. IEEE Acoustics, Speech, and Signal Proc. Conf., Tulsa, Oklahoma, 320-325 (1978)

Fastl, H., "Temporal Masking Effects: I. Broad-Band Noise Masker," Acoustica 35, 287-302 (1976).

Gallagher, R. G., Information Theory and Reliable Communication, 442-500 (Wiley, New York, 1968).

Green, D. M., "Additivity of Masking," J. Acoust. Soc. Am. 41, 1517-1525 (1967).

Jayant, N. S. and Rabiner, L. R., "The Application of Dither to the Quantization of Speech Signals," Bell System Tech. J. 51, 1293-1304 (1972).

Johnson, D. H., "The Response of Single Auditory-Nerve Fibers in the Cat to Single Tones: Synchrony and Average Discharge Rate." M.I.T. Ph.D thesis (1974).

Kiang, N. Y. S.; Watanabe, T.; Thomas, E. C.; and Clark, L. F., Discharge Patterns of Single Fibers in the Cat's Auditory Nerve, Research Monograph 35, (The M.I.T. Press, Cambridge, Mass., 1965).

Myers, J. P. and Fienberg, A., "High-Quality Professional Recording Using New Digital Techniques," J. Audio Eng. Soc. 20, 622-628 (1972).

Oppenheim, A. V. and Schafer, R. W., Digital Signal Processing (Prentice Hall, Englewood Cliffs, N.J., 1975).

Reed, C. M. and Bilger, R. C., "A Comparative Study of S/No and E/No," J. Acoust. Soc. Am. 53, 1039-1044 (1973).

Rhode, W. S., "Observations of the Vibration of the Basilar Membrane in Squirrel Monkeys Using the Mossbauer Technique," J. Acoust. Soc. Am. 48, 1218-1231 (1970).

Roberts, L. G., "Picture Coding Using Pseudo-Random Noise," IRE Information Theory Trans. IT-8, 145-152 (1962).

Robinson, C. E. and Pollack, I., "Interaction Between Forward and Backward Masking: A Measure of the Integrating Period of the Auditory System," J. Acoust. Soc. Am. 53, 1313-1316 (1973).

Sato, N., "PCM Recorder – A New Type of Audio Magnetic Tape Recorder," J. Audio Eng. Soc. 21, 542-548 (1973).

Schroeder, M. R., "Models of Hearing," Proc. IEEE, 63, 1332-1351 (1975).

Siebert, W. M., "*Frequency Discrimination in the Auditory System: Place or Periodicity Mechanisms?*," Proc. IEEE 58, 723-730 (1970).

Tribolet, J. M.; Noll, P.; McDermott, B. J.; and Crochiere, R. E., "Complexity and Quality of Speech Waveform Coders," Proc. IEEE Acoustics, Speech, and Signal Proc. Conf., Tulsa, Oklahoma, 586-590 (1978).

Viswanathan, R.; Makhoul, J.; and Huggins, A. W. F., "Speech Compression and Evaluation," Bolt, Beranek, and Newman, Inc., BBN Report 3794, 78-134 (1978).

Willcocks, M., "A Review of Digital Audio Techniques," J. Aud. Eng. Soc. 26, 56-64 (1978).

Zwicker, E., "Procedure for Calculating the Loudness of Temporally Variable Sounds," J. Acoust. Soc. Am. 62, 675-682 (1977).

APPENDIX I
ANALYSIS OF QUADRATURE MIRROR FILTERING

A. BASIC TECHNIQUE

The basic quadrature mirror-filter technique is designed to divide the digital frequency spectrum into two equal parts. Each band is sampled at half the rate of the original signal for quantization, coding, and transmission. The signals are then resampled back up to the original rate and filtered again before summing. Since the filters are not ideal filters, there is some aliasing after the down-sampling. Because of the special relationship of the filters, when the two bands are summed, the components due to the aliasing of one band cancel the components due to aliasing of the other band. Thus, there is no aliasing in the resultant signal.

The structure of the basic filter block was shown in Fig. IV-4. After down-sampling, the z-transforms of the signals are related by Eq. (I-2):

$$X_1(z) = H_1(z)\, X(z) \tag{I-1a}$$

$$X_2(z) = H_2(z)\, X(z) \tag{I-1b}$$

$$\begin{aligned} Y_1(z) &= \frac{1}{2}\, [X_1(z^{1/2}) + X_1(-z^{1/2})] \\ &= \frac{1}{2}\, [H_1(z^{1/2})\, X(z^{1/2}) + H_1(-z^{1/2})\, X(-z^{1/2})] \end{aligned} \tag{I-2a}$$

$$\begin{aligned} Y_2(z) &= \frac{1}{2}\, [X_2(z^{1/2}) + X_2(-z^{1/2})] \\ &= \frac{1}{2}\, [H_2(z^{1/2})\, X(z^{1/2}) + H_2(-z^{1/2})\, X(-z^{1/2})] \quad . \end{aligned} \tag{I-2b}$$

The second terms of Eqs. (I-2a and b) represent the aliasing due to under-sampling. The resampled process again scales the frequency axis:

$$\begin{aligned} U_1(z) &= Y_1(z^2) \\ &= \frac{1}{2}\, [H_1(z)\, X(z) + H_1(-z)\, X(-z)] \end{aligned} \tag{I-3a}$$

$$\begin{aligned} U_2(z) &= Y_2(z^2) \\ &= \frac{1}{2}\, [H_2(z)\, X(z) + H_2(-z)\, X(-z)] \quad . \end{aligned} \tag{I-3b}$$

Filtering again and summing:

$$\begin{aligned} T_1(z) &= K_1(z)\, U_1(z) \\ &= \frac{1}{2}\, [K_1(z)\, H_1(z)\, X(z) + K_1(z)\, H_1(-z)\, X(-z)] \end{aligned} \tag{I-4a}$$

$$\begin{aligned} T_2(z) &= K_2(z)\, U_2(z) \\ &= \frac{1}{2}\, [K_2(z)\, H_2(z)\, X(z) + K_2(z)\, H_2(-z)\, X(-z)] \end{aligned} \tag{I-4b}$$

$$S(z) = T_1(z) + T_2(z)$$

$$= \frac{1}{2} [H_1(z) K_1(z) + H_2(z) K_2(z)] X(z)$$

$$+ \frac{1}{2} [H_1(-z) K_1(z) + H_2(-z) K_2(z)] X(-z) \quad . \qquad \text{(I-5)}$$

The first term of Eq. (I-5) represents the linear filtered components of the signal. If the downsampling was removed from the process, this term would remain (scaled by a factor of 2) while the second term, due to aliasing, would vanish. If the filters were ideal nonoverlapping lowpass and highpass filters, the aliasing term would disappear then also.

When realizable filters are used, the aliasing terms can still be made to cancel. One simple solution is to let the filters be related as follows:

$$h_1(n) = h(n) \qquad \text{(I-6a)}$$

$$h_2(n) = (-1)^n h(n) \qquad \text{(I-6b)}$$

$$k_1(n) = h(n) \qquad \text{(I-6c)}$$

$$k_2(n) = -(-1)^n h(n) \quad . \qquad \text{(I-6d)}$$

The relations of the z-transforms of the filters may now be substituted in Eq. (I-5):

$$H_1(z) = H(z) \qquad \text{(I-7a)}$$

$$H_2(z) = H(-z) \qquad \text{(I-7b)}$$

$$K_1(z) = H(z) \qquad \text{(I-7c)}$$

$$K_2(z) = -H(-z) \qquad \text{(I-7d)}$$

$$S(z) = \frac{1}{2} [H(z) H(z) - H(-z) H(-z)] X(z)$$

$$+ \frac{1}{2} [H(-z) H(z) - H(-z) H(z)] X(-z)$$

$$= \frac{1}{2} [H^2(z) - H^2(-z)] X(z) \quad . \qquad \text{(I-8)}$$

The term that is left in Eq. (I-8) represents the reconstructed signal after being processed by the linear filters present in the system. Note that there have been no restrictions up to this point on the basic filter, h(n), only on the relationships of the other filters to h(n). Constraints on the filter so that the reconstructed signal will be identical to the original will be discussed in Section I-B. If the filter is a good approximation to the ideal half-band lowpass filter, the band will be divided into two equal bandwidth frequency bands.

Summarizing, if the relations of Eqs. (I-6) and (I-7) are used, the signal components due to aliasing are cancelled, leaving only linearly filtered signal components in the output. If the filter is chosen such that:

$$\left|\frac{1}{2} [H^2(z) - H^2(-z)]\right| = 1 \qquad \text{(I-9)}$$

and the phase is linear, then the system will be an identity system except for a linear phase component; i.e., a delay.

The quadrature mirror-filter system can be simply extended to divide the frequency range into more than two bands. A system that implements a decomposition into four bands was shown in Fig. IV-5. If the conditions on the filters are met such that the basic two-band system is an identity system, then application of that processing in the middle of another system will not affect the overall system output. It is clear that in that case the output in Fig. IV-5 will be identical to the input.

Section I-B will show that it is not practical to use filters designed such that the system is an identity system. It is important, therefore, to analyze the decomposition into four bands for filters that are related only by Eqs. (I-6) and (I-7), the relation that guarantees that the aliasing components in the basic two-channel decomposition cancel; i.e., that it acts like a linear filter.

The outputs of the inner decomposition are related to their inputs by the system function, G(z), of the two-band decomposition:

$$G(z) = \frac{1}{2}\,[H^2(z) - H^2(-z)] \tag{I-10}$$

$$Y'_1(z) = G(z)\,Y_1(z) \tag{I-11a}$$

$$Y'_2(z) = G(z)\,Y_2(z) \quad . \tag{I-11b}$$

The analysis continues by taking the equations derived earlier in the Section and modifying to account for the additional processing. The (I-3) equations now become:

$$\begin{aligned} U'_1(z) &= Y'_1(z^2) \\ &= G(z^2)\,Y_1(z^2) \\ &= G(z^2)\,\frac{1}{2}\,[H_1(z)\,X(z) + H_1(-z)\,X(-z)] \end{aligned} \tag{I-12a}$$

$$\begin{aligned} U'_2(z) &= Y'_2(z^2) \\ &= G(z^2)\,Y_2(z^2) \\ &= G(z^2)\,\frac{1}{2}\,[H_2(z)\,X(z) + H_2(-z)\,X(-z)] \quad . \end{aligned} \tag{I-12b}$$

Filtering and summing:

$$\begin{aligned} S'(z) &= G(z^2)\,\frac{1}{2}\,[H_1(z)\,K_1(z) + H_2(z)\,K_2(z)]\,X(z) \\ &\qquad + G(z^2)\,\frac{1}{2}\,[H_1(-z)\,K_1(z) + H_2(-z)\,K_2(z)]\,X(-z) \\ &= G(z^2)\,\frac{1}{2}\,[H^2(z) - H^2(-z)]\,X(z) \\ &= G(z^2)\,G(z)\,X(z) \quad . \end{aligned} \tag{I-13}$$

Again, the aliasing components have vanished leaving a linear filtered system. The new system function is:

$$G'(z) = \frac{S'(z)}{X(z)} = G(z^2)\,G(z) \quad . \tag{I-14}$$

This equation also validates the earlier conjecture that if G(z) is an identity system, then G'(z) will be also.

When error is introduced to a channel through quantization of the signal, that noise is filtered and will result in noise at the output. The basic two-channel quadrature mirror-filter system with quantization, a nonlinear function modeled as an additive error signal, was shown in Fig. IV-6. Since resampling, filtering, and summation are linear functions, this additive noise term will be an additive processed noise term at the output as shown in Eq. (I-15):

$$S(z) = G(z)\,X(z) + K_1(z)\,E_1(z^2) + K_2(z)\,E_2(z^2) \quad . \tag{I-15}$$

Resampling affects the error signal by scaling the frequency axis. If the additive error is wideband, as can normally be assumed, this will not change its spectral extent. The error from quantization of a band in the two-band system will result in a wideband noise that is filtered once in that channel.

The multiband system is more complex as was illustrated by the four-band system in Fig. IV-7 and analyzed below:

$$\begin{aligned} S(z) &= G(z^2)\,G(z)\,X(z) \\ &\quad + [K_1(z^2)\,E_1(z^4) + K_2(z^2)\,E_2(z^4)]\,K_1(z) \\ &\quad + [K_1(z^2)\,E_4(z^4) + K_2(z^2)\,E_3(z^4)]\,K_2(z) \\ &= G(z^2)\,G(z)\,X(z) \\ &\quad + K_1(z)\,K_1(z^2)\,E_1(z^4) + K_1(z)\,K_2(z^2)\,E_2(z^4) \\ &\quad + K_2(z)\,K_2(z^2)\,E_3(z^4) + K_2(z)\,K_1(z^2)\,E_4(z^4) \quad . \end{aligned} \tag{I-16}$$

Resampling of the filtered signal of each channel will change the effective filter frequency response by warping of the frequency axis. The spectra of these error signals was shown in Fig. IV-8. It was noted then that this warping will sharpen the filtering on some, but not all of the frequency channels.

B. DESIGN OF THE QUADRATURE MIRROR FILTERS FOR PERFECT RECONSTRUCTION

In Section A, it was shown that the frequency band could be subdivided into two equal-bandwidth mirror-image bands using realizable, overlapping filters. Each of these bands can be resampled at half the original sampling rate and still allow later recovery of the original signal if the filters obey certain simple relationships. The aliasing present after down-sampling is cancelled when the bands are summed when the four filters in the system are designed from one arbitrary prototype filter, h(n). Restrictions on this filter are placed solely so that the linear filtering and summation operations result in an acceptable system. This Section is concerned with the design of the filter.

The output of the system was described by Eq. (I-8). It is desired that the system be an identity system (except for a linear phase term necessitated by the use of causal, realizable filters). Since the nonlinear aliasing terms have vanished, an impulse response of the system and its transform, the system function, can be defined:

$$G(z) = \frac{S(z)}{X(z)} = \frac{1}{2}\,[H^2(z) - H^2(-z)] \tag{I-17}$$

$$G(z) = z^{-(N-1)} \quad \text{is desired} \quad . \tag{I-18}$$

Evaluation of the system function on the unit circle in the z-plane is the Fourier transform of the impulse response:

$$G(\omega) = \frac{1}{2}\,[H^2(\omega) - H^2(\omega + \pi)] \tag{I-19}$$

$$G(\omega) = \exp[-j\omega(N-1)] \quad \text{is desired} \quad . \tag{I-20}$$

The desired impulse response is just a delayed impulse. In order to design h(n), it is necessary to see what constraints this imposes.

Let the filter $H(\omega)$ be a real, causal, symmetric, FIR filter of length N. The linear phase term may be factored out of the Fourier transform leaving a real and even filter function, denoted as $H'(\omega)$. This is then substituted in Eq. (I-17):

$$H(\omega) = H'(\omega)\exp[-j\omega(N-1)/2] \tag{I-21}$$

$$\begin{aligned} G(\omega) &= \frac{1}{2}\,\{H'^2(\omega)\exp[-j\omega(N-1)] - H'^2(\omega+\pi)\exp[-j(\omega+\pi)(N-1)]\} \\ &= \frac{1}{2}\exp[-j\omega(N-1)]\,[H'^2(\omega) - (-1)^{N-1}H'^2(\omega+\pi)] \\ &= \frac{1}{2}\exp[-j\omega(N-1)]\,[H'^2(\omega) + (-1)^{N}H'^2(\omega+\pi)] \quad . \end{aligned} \tag{I-22}$$

Since it was assumed that $H(\omega)$ is a real, symmetric FIR filter, $H'(\omega)$ is a real and even filter. Thus,

$$H'^2(\omega) = H'^2(\omega+\pi) \qquad \text{evaluated at } \omega = \frac{\pi}{2} \quad . \tag{I-23}$$

Unless the length of the filter, N, is even, the system response, $G(\omega)$, must have a zero at that frequency. (Modification of the system to permit use of odd filter lengths will be discussed later in this Section.)

G(z) may now be found in terms of the transform of h(n) using Eq. (I-17) and letting the length of the filter be even:

$$H(z) = h_0 + h_1 z^{-1} + \ldots + h_{N-1} z^{-(N-1)}$$

where

$$h_n = h_{N-1-n} \quad \text{and} \quad h_0 = h_{N-1} \neq 0 \quad . \tag{I-24}$$

$$H^2(z) = a_0 + a_1 z^{-1} + \ldots + a_{2N-2} z^{-(2N-2)} \tag{I-25a}$$

$$H^2(-z) = a_0 - a_1 z^{-1} + \ldots + a_{2N-2} z^{-(2N-2)} \tag{I-25b}$$

$$G(z) = a_1 z^{-1} + a_3 z^{-3} + \ldots + a_{2N-3} z^{-(2N-3)} \tag{I-25c}$$

where

$$a_n = h_0 h_n + h_1 h_{n-1} + \ldots + h_n h_0 \quad . \tag{I-25d}$$

In the summation to form G(z), all of the even-subscripted terms have dropped out. All of the odd-subscripted terms other than $n = (N - 1)$ must be set to zero and the equations solved using Eq. (I-25d):

$$a_1 = 2h_0 h_1 = 0$$

since

$$h_0 \neq 0 \quad , \quad \text{then } h_1 = 0 \quad . \tag{I-26a}$$

$$a_3 = 2\,[h_0 h_3 + h_1 h_2] = 0$$

since

$$h_0 \neq 0 \quad \text{and} \quad h_1 = 0 \quad , \quad \text{then } h_3 = 0 \quad . \tag{I-26b}$$

By induction for all n odd, $n \neq N - 1$:

$$a_n = 2\,[h_0 h_n + h_1 h_{n-1} + \ldots + h_{(n-1)/2} h_{(n+1)/2}] = 0$$

since

$$h_0 \neq 0 \quad , \quad h_1 = h_3 = \ldots = h_{n-2} = 0 \quad , \quad \text{then } h_n = 0 \quad . \tag{I-26c}$$

For $n = N - 1$:

$$a_{N-1} = 2\,[h_0 h_n + h_1 h_{n-1} + \ldots + h_{(n-1)/2} h_{(n+1)/2}]$$

$$= 2h_0 h_{N-1} = 1$$

thus

$$h_0 = h_{N-1} = 2^{-1/2} \quad . \tag{I-26d}$$

$$H(z) = 2^{-1/2} + 2^{-1/2}\, z^{N-1} \quad . \tag{I-27}$$

All of the odd-subscripted terms of h(n) must be zero except for $n = (N - 1)$. By the symmetric nature of the filter, all of the even-subscripted terms except $n = 0$ must also equal zero. Hence, the only filters that can result in the desired response are identically zero except at each of the end points. This represents a class of filters with zeros at each of the $N - 1$ roots of -1. Unfortunately, this class of filters does not include any suitable half-band lowpass filters for the system.

It is interesting to relate this result to the use of the discrete Fourier transform (DFT) to effect a similar transformation of domains. The DFT, an invertible function, transforms N points in the time domain to N points in the frequency domain. The DFT may be implemented by down-sampling the outputs of a set of N linear filters.[43] The impulse responses of the filters are:

$$h_k(n) = \exp[-j2\pi kn/N] \quad , \quad 0 \leq n \leq N-1$$

$$0 \quad , \quad \text{otherwise}$$

$$\text{for filter } k \quad , \quad 0 \leq k \leq N-1 \quad . \qquad \text{(I-28)}$$

A set of N DFT output samples is obtained, one at the output of each filter for every N input samples by sampling each filter output once every N samples. Thus, the DFT can be implemented as a filter bank of N channels with outputs that can be resampled each at 1/N the original rate. The sum of the sampling rates of the channels is the original sampling rate. The DFT and the inverse discrete Fourier transform (IDFT) form an identity system in the same form as is desired for the digital encoder. From Eq. (I-28), the impulse responses of the filters used in the 2-point DFT are:

$$h_0(n) = 1 \quad , \quad n = 0 \quad , \quad 1$$

$$= 0 \quad , \quad \text{otherwise}$$

$$h_1(n) = 1 \quad , \quad n = 0$$

$$= -1 \quad , \quad n = 1$$

$$= 0 \quad , \quad \text{otherwise} \quad . \qquad \text{(I-29)}$$

Comparing the DFT filters with the filters used in the 2-channel quadrature mirror-filter scheme, the DFT filters are (to within a constant which is in the IDFT filters) the prototype filter and its mirror filter needed for an identity system. Thus, the 2-point DFT system is an identity-quadrature, mirror-filter system.

As was determined in Section A, the use of linear phase FIR filters of odd length results in a zero of the system response at $\omega = \pi/2$. The basic quadrature mirror-filter relationships can be modified so that odd-length filters may be acceptable. By changing the relationships of the filters, the system function is altered. The modification requires the insertion of a delay in one channel when filtering and a corresponding delay in the other channel when refiltering. The aliasing components are still cancelled. The new relations are described in Eq. (I-30) along with their z-transforms:

$$h_1(n) = h(n) \quad \longleftrightarrow \quad H_1(z) = H(z)$$

$$h_2(n) = (-1)^{n-1}\, h(n-1) \quad \longleftrightarrow \quad H_2(z) = z^{-1}\, H(-z)$$

$$k_1(n) = h(n-1) \quad \longleftrightarrow \quad K_1(z) = z^{-1}\, H(z)$$

$$k_2(n) = (-1)^n\, h(n) \quad \longleftrightarrow \quad K_2(z) = H(-z) \quad . \qquad \text{(I-30)}$$

Substituting these new relations in the equation of the output of the basic block scheme of Fig. IV-4, Eq. (I-5):

$$S(z) = \frac{1}{2}\,[H_1(z)\,K_1(z) + H_2(z)\,K_2(z)]\,X(z) + \frac{1}{2}\,[H_1(-z)\,K_1(z) + H_2(-z)\,K_2(z)]\,X(-z)$$

$$= \frac{1}{2}\,[H(z)\,z^{-1}\,H(z) + z^{-1}\,H(-z)\,H(-z)]\,X(z) + \frac{1}{2}\,[H(-z)\,z^{-1}\,H(z) + (-z)^{-1}\,H(z)\,H(-z)]\,X(-z)$$

$$= \frac{1}{2}\,z^{-1}\,[H^2(z) + H^2(-z)]\,X(z) \quad . \qquad \text{(I-31)}$$

As before, the aliasing terms have vanished leaving the linear filtered terms only. Proceeding by defining an impulse response and its z-transform:

$$G(z) = \frac{S(z)}{X(z)} = \frac{1}{2} z^{-1} [H^2(z) + H^2(-z)] \quad . \tag{I-32}$$

This equation differs from the system function of the original, unmodified filter relations, Eq. (I-17), only by a sign change; i.e., the summation of the terms rather than the difference, and a delay of one sample.

Assume now that the filter, h(n), is a symmetric FIR filter of order N. The linear phase term may be factored out of the transform. Evaluating the system function on the unit circle as in Eq. (I-22):

$$H(\omega) = H'(\omega) \exp[-j\omega(N-1)/2]$$

$$G(\omega) = \frac{1}{2} \exp[-j\omega] \{H'^2(\omega) \exp[-j\omega(N-1)] + H'^2(\omega+\pi) \exp[-j(\omega+\pi)(N-1)]\}$$

$$= \frac{1}{2} \exp[-j\omega N] [H'^2(\omega) + (-1)^{N-1} H'^2(\omega+\pi)] \quad . \tag{I-33}$$

Since

$$H'^2(\omega) = H'^2(\omega+\pi) \quad \text{at } \omega = \frac{\pi}{2} \quad ,$$

there will be a zero of the system function at that frequency unless the order of the filter, N, is odd.

Perhaps it is now possible to find a suitable filter of odd length that will yield an identity system:

$$H(z) = h_0 + h_1 z^{-1} + \ldots + h_{N-1} z^{-(N-1)}$$

where

$$h_k = h_{N-1-k} \quad \text{and} \quad h_0 = h_{N-1} \neq 0 \quad . \tag{I-34}$$

$$H^2(z) = a_0 + a_1 z^{-1} + \ldots + a_{2N-2} z^{-(2N-2)}$$

$$H^2(-z) = a_0 - a_1 z^{-1} + \ldots + a_{2N-2} z^{-(2N-2)}$$

$$G(z) = \frac{1}{2} z^{-1} [a_0 + a_2 z^{-2} + \ldots + a_{2N-2} z^{-(2N-2)}]$$

where

$$a_n = h_0 h_n + h_1 h_{n-1} + \ldots + h_n h_0 \quad . \tag{I-35}$$

But

$$a_0 = h_0^2 \neq 0 \quad \text{and} \quad a_{2N-2} = h_{N-1}^2 \neq 0 \quad .$$

To make G(z) as desired, all of the terms must vanish except for n = (N − 1). For h(n) to be a length N filter, h(0) and H(N − 1) can not be zero. Then, two of the terms above are nonzero. Thus, there are no filters of odd length that would result in an identity system.

C. UNEQUAL BANDWIDTH FILTER BANK DESIGN

The three-channel quadrature mirror-filter system shown in Fig. IV-9 will result in some aliasing. Analysis of the equations show the problems involved with the aliasing:

$$X_1(z) = H_1(z)\, X(z) \tag{I-36a}$$

$$X_2(z) = H_2(z)\, X(z) \tag{I-36b}$$

$$\begin{aligned} Y_1(z) &= \tfrac{1}{2}\left[X_1(z^{1/2}) + X_1(-z^{1/2})\right] \\ &= \tfrac{1}{2}\left[H_1(z^{1/2})\, X(z^{1/2}) + H_1(-z^{1/2})\, X(-z^{1/2})\right] \end{aligned} \tag{I-37a}$$

$$\begin{aligned} Y_2(z) &= \tfrac{1}{2}\left[X_2(z^{1/2}) + X_2(-z^{1/2})\right] \\ &= \tfrac{1}{2}\left[H_2(z^{1/2})\, X(z^{1/2}) + H_2(-z^{1/2})\, X(-z^{1/2})\right] \end{aligned} \tag{I-37b}$$

$$\begin{aligned} U'_1(z) &= Y'_1(z^2) \\ &= G(z^2)\, Y_1(z^2) \\ &= \tfrac{1}{2}\, G(z^2)\left[H_1(z)\, X(z) + H_1(-z)\, X(-z)\right] \end{aligned} \tag{I-38a}$$

$$\begin{aligned} U_2(z) &= Y_2(z^2) \\ &= \tfrac{1}{2}\left[H_2(z)\, X(z) + H_2(-z)\, X(-z)\right] \quad . \end{aligned} \tag{I-38b}$$

Filtering again and summing:

$$\begin{aligned} T'_1(z) &= K_1(z)\, U'_1(z) \\ &= \tfrac{1}{2}\, G(z^2)\left[K_1(z)\, H_1(z)\, X(z) + K_1(z)\, H_1(-z)\, X(-z)\right] \end{aligned} \tag{I-39a}$$

$$\begin{aligned} T_2(z) &= K_2(z)\, U_2(z) \\ &= \tfrac{1}{2}\left[K_2(z)\, H_2(z)\, X(z) + K_2(z)\, H_2(-z)\, X(-z)\right] \end{aligned} \tag{I-39b}$$

$$\begin{aligned} S(z) &= T'_1(z) + T_2(z) \\ &= \tfrac{1}{2}\left[H_1(z)\, K_1(z)\, G(z^2) + H_2(z)\, K_2(z)\right] X(z) \\ &\quad + \tfrac{1}{2}\left[H_1(-z)\, K_1(z)\, G(z^2) + H_2(-z)\, K_2(z)\right] X(-z) \\ &= \tfrac{1}{2}\left[H^2(z)\, G(z^2) - H^2(-z)\right] X(z) \\ &\quad + \tfrac{1}{2}\left[G(z^2) - 1\right]\left[H(z)\, H(-z)\right] X(-z) \quad . \end{aligned} \tag{I-40}$$

The second term in Eq. (I-40) represents the aliasing in the output. As stated in Chapter IV-4, the aliasing will only appear in the frequency overlap of the filter h(n) and its mirror filter. Since

$$G(z^2) \approx 1$$

in the overlap region, the aliasing will be very small.

APPENDIX II

STATISTICAL ANALYSIS OF EXPERIMENTAL PROCEDURE

Statistical analysis of the experiments to evaluate the performance of the digital encoding system gives further insight into the problem of meaningful evaluation and comparison of speech and audio processing systems. The differential threshold of encoding degradation was defined in Chapter V as the degradation that yields a 75-percent probability of a correct response in the 2I2AFC comparisons. This JND threshold is a mean probability of a correct response averaged over an ensemble of the entire population of prospective users. The ability of a subject to detect differences is not necessarily the same as other subjects and can be modeled as a sample of a random variable. This random variable, the probability of a correct response, is assumed to have a Gaussian distribution. The standard deviation of the distribution is a measure of how much the detection ability of each person varies from the average. The experimental analysis problem is twofold:

(a) Estimate the probability of a correct response for each subject via their responses on several trials

(b) Estimate the average probability of the ensemble and the variance of the density function from the estimates of the individual subject probability.

For a given subject, subject m, estimate the probability of a correct response. Let x be the random variable of the response to a trial, 0 if not correct, 1 if correct,

$$\text{and let } \hat{P}_m = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{be the estimate of } P_m \ . \tag{II-1}$$

where

$x_i = 0$ if not correct on the i^{th} trial which occurs with probability $(1 - P_m)$

$x_i = 1$ if correct on the i^{th} trial which occurs with probability P_m

N = number of trials

$P_m = E[x]$ = Probability of a correct response for subject m. Expectation operator defined over the ensemble of trials for a particular subject.

$$E[\hat{P}_m] = P_m \tag{II-2}$$

$$\mathrm{Var}[\hat{P}_m] = \frac{P_m(1 - P_m)}{N} \tag{II-3}$$

This estimator is an unbiased and efficient estimator. If the number of trials, N, is large, the distribution of the estimate can be approximated well by a normal density function with mean and variance given above. Evaluated for 48 trials and a subject probability of 75 percent, the standard deviation of the estimator is 6.25 percent.

Using these estimates of the statistics for each subject it is now possible to estimate the density function over the population of subjects. Let P be the random variable with samples estimated above in Eq. (II-1) and a density which is assumed to be normal. The number of subjects is M. Assume that:

$$P = N(\mu, \sigma^2) \quad . \tag{II-4}$$

Let:

$$\hat{\mu} = \frac{1}{M} \sum_m \hat{P}_m \quad \text{be estimate of the mean.} \tag{II-5}$$

$$\hat{\sigma}^2 = \frac{1}{M-1} \sum_m (\hat{P}_m - \hat{\mu})^2 \quad \text{be estimate of the variance.} \tag{II-6}$$

The distribution of the estimator of the mean is described by Eqs. (II-7) and (II-8). The second term in Eq. (II-8) is due to estimating μ by the estimates of the individual subject probabilities rather than the exact probabilities:

$$E[\hat{\mu}] = \mu \tag{II-7}$$

$$\begin{aligned} \text{Var}[\hat{\mu}] &= \frac{\sigma^2}{M} + \frac{1}{M^2} \sum_{m=1}^{M} \frac{P_m(1-P_m)}{N} \\ &\approx \frac{\hat{\sigma}^2}{M} + \frac{\hat{\mu}(1-\hat{\mu})}{MN} \quad . \end{aligned} \tag{II-8}$$

The results of the experiments presented in Table V-1 can now be analyzed. Note that it is actually the estimates of the individual sample probabilities that are used. The experiment with five subjects using the quantization bit distribution parameters denoted as System E is:

$$\hat{\mu} = 0.722 \tag{II-9}$$

$$\hat{\sigma}^2 = (0.0559)^2 \tag{II-10}$$

$$\text{Var}[\hat{\mu}] \approx (0.0250)^2 + (0.0289)^2 = (0.0382)^2 \tag{II-11}$$

For System E, the estimated mean is 72.2 percent. The distribution of this estimate of the mean is approximately Gaussian with mean 72.2 percent and standard deviation of 2.5 percent. The estimate of the variance for System E gives a standard deviation of 5.59 percent. Assuming the normal distribution, the above estimates can be used to make the following claims for System E: 50 percent of the population have a probability of a correct response on a trial of less than 72.2 percent; 69 percent have a probability of less than 75 percent; 84 percent have a probability of less than 78 percent; 93 percent have a probability of less than 81 percent; and 98 percent have a probability of less than 84 percent. Since the criterion for the JND was set at 75 percent a priori, it is concluded that over the population of subjects, 31 percent will be able to discern a difference between unprocessed speech and speech processed by System E in a test like the one performed here. Presentation of continuous speech, however, is equivalent to many trials. The probability of detection of the encoding degradation is larger than for a single sentence pair and is a function of the length of the presentation of continuous speech.

LIST OF ABBREVIATIONS

A/D	Analog to digital
ADM	Adaptive delta modulation
ADPCM	Adaptive differential pulse code modulation
APC	Adaptive predictive coding
APCM	Adaptive pulse code modulation
ATC	Adaptive transform coding
CCD	Charge-coupled device
D/A	Digital to analog
dB	Decibels
DFT	Discrete Fourier transform
DM	Delta modulation
FFT	Fast Fourier transform
FIR	Finite impulse response
IDFT	Inverse discrete Fourier transform
IIR	Infinite impulse response
JND	Just noticeable difference (differential threshold)
kbps	Kilobits per second
kHz	Kilohertz
PCM	Pulse code modulation
RMS	Root mean square
SNR	Signal-to-noise ratio
SPL	Sound pressure level, re 0.0002 dyne/sq cm
2I2AFC	Two-interval, two-alternative forced choice

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

19 REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 18 ESD-TR-79-154	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (*and Subtitle*) 6 Digital Encoding of Speech and Audio Signals Based on the Perceptual Requirements of the Auditory System,		5. TYPE OF REPORT & PERIOD COVERED 9 Technical Report, 6. PERFORMING ORG. REPORT NUMBER Technical Report 535
7. AUTHOR(*s*) 10 Michael A. Krasner		8. CONTRACT OR GRANT NUMBER(*s*) 15 F19628-78-C-0002, WARPA Order-2006
9. PERFORMING ORGANIZATION NAME AND ADDRESS Lincoln Laboratory, M.I.T. P.O. Box 73 Lexington, MA 02173		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS ARPA Order 2006 Program Element No. 62706E 16 Project No. 9P10
11. CONTROLLING OFFICE NAME AND ADDRESS Defense Advanced Research Projects Agency 1400 Wilson Boulevard Arlington, VA 22209		12. REPORT DATE 11 18 Jun[redacted]979 13. NUMBER OF PAGES 76
14. MONITORING AGENCY NAME & ADDRESS (*if different from Controlling Office*) Electronic Systems Division Hanscom AFB Bedford, MA 01731 12 73		15. SECURITY CLASS. (*of this report*) Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (*of this Report*)

Approved for public release; distribution unlimited.

14 TR-535

17. DISTRIBUTION STATEMENT (*of the abstract entered in Block 20, if different from Report*)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (*Continue on reverse side if necessary and identify by block number*)

digital encoding system	quadrature mirror-filter technique	adaptive quantization
auditory system	speech and audio signals	predictive quantization
psychoacoustic experiments	instantaneous quantization	

20. ABSTRACT (*Continue on reverse side if necessary and identify by block number*)

The development of a digital encoding system for speech and audio signals is described. The system is designed to exploit the limited detection ability of the auditory system. Existing digital encoders are examined. Relevant psychoacoustic experiments are reviewed. Where the literature is lacking, a simple masking experiment is performed and the results reported. The design of the encoding system and specifications of system parameters are then developed from the perceptual requirements and digital signal processing techniques.

The encoder is a multi-channel system, each channel approximately of critical bandwidth. The input signal is filtered via the quadrature mirror-filter technique. An extensive development of this technique is presented. Channels are quantized with an adaptive PCM scheme.

The encoder is evaluated for speech and audio signal inputs. For 4.1-kHz bandwidth speech, the differential threshold of encoding degradation occurs at a bit rate of 34.4 kbps. At 16 kbps, the encoder produces toll-quality speech output. Audio signals of 15-kHz bandwidth can be encoded at 123.8 kbps without audible degradation.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

207 650 mt